



**Universität  
Basel**

Wirtschaftswissenschaftliche  
Fakultät



---

June 2018

# **Identification of causal mechanisms based on between-subject double randomization designs**

WWZ Working Paper 2018/19

Renate Strobl, Conny Wunsch

---

A publication of the Center of Business and Economics (WWZ), University of Basel.

© WWZ 2018 and the authors. Reproduction for other purposes than the personal use needs the permission of the authors.

---

Universität Basel  
Peter Merian-Weg 6  
4052 Basel, Switzerland  
wwz.unibas.ch

**Corresponding Author:**  
Dr. Renate Strobl  
Tel: +41 (0) 61 207 23 96  
Mail: [renate.strobl@unibas.ch](mailto:renate.strobl@unibas.ch)

# Identification of causal mechanisms based on between-subject double randomization designs

Conny Wunsch\* and Renate Strobl†‡

June 30, 2018

## Abstract

Understanding the mechanisms through which treatment effects come about is crucial for designing effective interventions. The identification of such causal mechanisms is challenging and typically requires strong assumptions. This paper discusses identification and estimation of natural direct and indirect effects in so-called double randomization designs that combine two experiments. The first and main experiment randomizes the treatment and measures its effect on the mediator and the outcome of interest. A second auxiliary experiment randomizes the mediator of interest and measures its effect on the outcome. We show that such designs allow for identification based on an assumption that is weaker than the assumption of sequential ignorability that is typically made in the literature. It allows for unobserved confounders that do not cause heterogeneous mediator effects. We demonstrate estimation of direct and indirect effects based on different identification strategies that we compare to our approach using data from a laboratory experiment we conducted in Kenya.

**Keywords:** Direct and indirect effects, causal inference, mediation analysis, identification

**JEL classification:** C31, D64

---

\*University of Basel, CEPR, CESifo, IZA, conny.wunsch@unibas.ch.

†University of Basel, rena.te.strobl@unibas.ch.

‡We thank the team of the Busara Center of Behavioral Economics in Nairobi, in particular Channing Jang, Georgina Mburu and all involved research assistants for helping to realize this project and their support. Moreover, we thank Annabelle Dörr, Martin Huber, Jeffrey A. Smith, Andreas Steinmayr, Anthony Strittmatter and seminar participants at the Universities of Basel, Lucerne, Neuchatel, Mannheim and Munich for helpful discussions and comments.

# 1 Introduction

Uncovering causal mechanisms through which some policy, intervention, or treatment affects some outcome of interest is crucial for understanding how causal effects come about. This, in turn, is a prerequisite for designing effective policy interventions. Causal mediation analysis (Robins and Greenland, 1992; Pearl, 2001; Robins, 2003) aims at identifying causal mechanisms, i.e. at disentangling the natural direct effect of the treatment on the outcome from indirect effects operating through one or more intermediate outcomes, which are also called mediators. The identification of causal mechanisms is much more challenging than identification of the overall causal effect of the treatment even if the latter has been randomized. The reason is that we need to separately identify the effect of the mediator on the outcome for those who endogenously respond to the treatment in terms of the mediator, which is a non-random subsample of the population of interest. The typical solution is to assume sequential ignorability of the mediator which requires the researcher to observe all confounders that jointly determine the mediator and the outcome of interest conditional on the treatment (Petersen et al., 2006; Flores and Flores-Lagunes, 2009; VanderWeele, 2009; Hong, 2010; Imai et al., 2010; Tchetgen Tchetgen and Shpitser, 2012; Zheng and van der Laan, 2012; Imai et al., 2013; Huber, 2014). Recently, some alternatives have been proposed in the literature that either require at least one instrumental variable (Imai et al., 2013; Yamamoto, 2014; Park, 2015; Frölich and Huber, 2017) or rely on common trend assumptions imposed on certain subpopulations that require multiple observations over time (Deuchert et al., 2016) and may only identify the effects of interest for certain sub-populations.

As an alternative, the literature proposes to combine the main experiment that randomizes the treatment and measures its effect on the mediator and the outcome of interest with an auxiliary experiment that randomizes the mediator and/or the treatment conditional on the mediator in order to identify natural direct and indirect effects. Imai et al. (2013) propose a so-called parallel design where the auxiliary experiment randomly assigns subjects to different values of the mediator and then randomizes the treatment conditional on the value of the mediator.<sup>1</sup> They show that in this case the natural direct and indirect effects are identified under the assumption that the treatment effect does not depend on the value of the mediator. As Imai et al. (2013) point out, this is also a strong assumption that might be just as hard to justify as sequential ignorability. Moreover, this identification result does not extend to designs where the auxiliary experiment randomizes the mediator for one particular treatment state or independent of the treatment, which have the advantage that they are relatively easy to implement. Examples of such designs have been discussed by Pirlott and MacKinnon (2016) but without showing formal identification results. It is intuitive, though, and has been pointed out by Imai et al. (2013), that designs that allow identifying the average effect of the mediator on the outcome by randomizing the

---

<sup>1</sup>They also propose so-called encouragement designs where auxiliary experiment generates an instrumental variable for the mediator.

mediator should be able to identify the direct and indirect effects under the assumption that the mediator effect is homogeneous. The reason is that in this case the effect of the mediator for those who endogenously respond to the treatment in terms of the mediator is the same as the average mediator effect. However, effect homogeneity is a strong assumption that is likely to be violated in many applications, which strongly limits the usefulness of such designs for identifying direct and indirect effects.

In this paper, we discuss identification in different double randomization designs where the auxiliary experiment randomizes the mediator and, hence, identifies the average mediator effect. We focus on between-subject designs because they allow studying actual behavior in the field. In contrast, within-subject designs limit applications to either hypothetical or sequential choices and they require that the order of the experiments does not affect behavior (Imai et al., 2013). As a first contribution, we formalize the intuitive idea that in designs that identify the average mediator effect, identification of the direct and indirect effects can be achieved by requiring homogeneity of the effect of the mediator on the outcome. We show, however, that in all but one design this assumption is not sufficient for identification. We further show that identification based on a homogeneous mediator effect results in a very simple estimand that combines two easy-to-compute effects, namely the average mediator effect that is identified from randomizing the mediator in the auxiliary experiment, and the average treatment effect on the mediator that is identified from the randomization of the treatment in the main experiment.

As the second and main contribution of this paper, we propose a new identification strategy that relaxes the assumption of a homogeneous mediator effect based on the idea that it is possible to handle effect heterogeneity caused by observed characteristics. We show that the resulting identifying assumption of conditional homogeneity of the mediator effect is weaker than the sequential ignorability assumption because it allows for some unobserved confounders, namely those that do not cause heterogeneous mediator effects. Because the mediator is randomized in the auxiliary experiment, extensive tests for heterogeneous effects, e.g. based on modern machine learning methods, can be used to select the covariates that render this assumption plausible. The strategy we propose allows re-weighting conditional mediator effects according to the distribution of the characteristics of the subjects that respond to the treatment in terms of the mediator. We propose a semi-parametric propensity score weighting estimator to estimate the direct and indirect effects based on this strategy that is  $\sqrt{N}$ -consistent and allows using bootstrapping methods for inference. The main difference to identification based on the sequential ignorability assumption is that we reweigh subjects to handle effect heterogeneity and not to handle selection bias.<sup>2</sup> Our strategy allows for identification without the possibly strong assumption that the treatment effect does not depend on the value of the mediator that is the core of the strategy proposed by Imai et al. (2013) for the parallel design. This implies that our strategy may still be used when this assumption is violated in parallel

---

<sup>2</sup>An approach that is similar in spirit has been proposed by Angrist and Fernandez-Val (2013) and Aronow and Carnegie (2013) to recover the average treatment effect (and other parameters) from the local average treatment effect in the context of identification of treatment effects based on instrumental variables in the presence of heterogeneous treatment effects. We discuss the difference to their case later in the paper.

designs. Moreover, the proposed strategy is applicable to all designs that identify the average mediator effect in the auxiliary experiment, which considerably improves the usefulness of between-subject double randomization designs for identifying direct and indirect effects. It implies that these designs can be used for identification based on an assumption that is weaker the sequential ignorability assumption.

As a third contribution, we show that for certain research questions it may be possible to implement an experimental design that allows identifying direct and indirect effects without having to impose additional assumptions. This refers to cases where the treatment of interest is endogenous choice of the mediator versus random assignment of the mediator. For example, consider caseworker assignment to a training program versus random assignment. A researcher may be interested in the question whether there is a value-added effect of caseworker assignment, e.g. a motivational effect, beyond the fact that program participants differ in both cases. In such designs, it is possible to elicit caseworker choices for all subjects in the experiment before actual assignment to the program. A random subgroup of subjects then receives the caseworker choice while the other group is randomly assigned to the program. As a result, we observe counterfactual caseworker choices for those randomly assigned to the program, which allows identifying direct and indirect effects without having to impose additional assumptions.

As the final contribution, we use such a design to demonstrate estimation of direct and indirect effects based on different identification strategies: sequential ignorability, global homogeneity of the mediator effect and the new approach we propose that is based on conditional homogeneity of the mediator effect. We use data from a laboratory experiment we conducted in Kenya that is analyzed in Wunsch and Strobl (2018). It investigates whether solidarity, which is a crucial base for informal insurance arrangements in developing countries, is sensitive to the extent to which individuals can influence their risk exposure. Our design measures subjects' monetary transfers to a partner who experiences a negative income shock both in a setting where participants could either deliberately choose or were randomly assigned to a safe or a risky project. Here, risk exposure is the mediator of interest and the treatment is free project choice (endogenous risk) versus random assignment of projects (exogenous risk). A unique feature of our experiment is that we directly measure preferred projects for all individuals, which allows us to identify the direct and indirect effect without having to impose additional assumptions. This provides us with an assumption-free benchmark that we can use to test the validity of the different identification strategies and to compare them with each other.

We find that validity of sequential ignorability cannot be rejected in our data, which implies that we managed to collect information on all relevant confounders within the experiment. This is an important input for non-experimental studies that analyze the relationship between risk-taking and solidarity. Furthermore, we find that identification based on the strong assumption of a homogeneous mediator effect is rejected by the data. However, when relaxing this assumption to conditional homogeneity of the mediator effect, we are able to match the results of the assumption-free benchmark. We also show that

this approach is indeed less demanding than identification based on sequential ignorability because we match the benchmark without controlling for the main driver of selection. As a further important result we find strong asymmetries in the treatment effects conditional on mediator value, which indicates that the assumption of no causal interaction between the treatment and the mediator, which is crucial for identification in the parallel design as proposed by Imai et al. (2013) but not for our approach, is violated in our application. This shows that this assumption is indeed a strong one and that the approach we propose is an important alternative that can be applied in parallel designs even if this assumption is violated, which can be tested in the data.

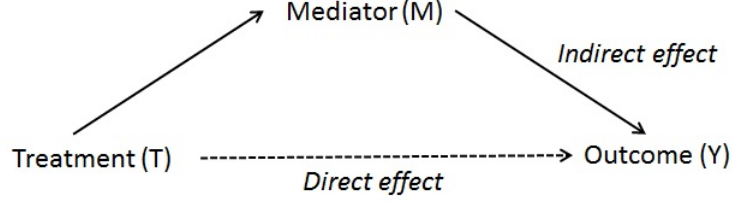
The remainder of the paper is organized as follows. The first part of the paper presents the methodological contribution. We start by describing our setup and the four double randomization designs we consider. Thereafter, we discuss identification starting with sequential ignorability. We then discuss approaches that exploit double randomization. We start by recapping identification in the parallel design as proposed by Imai et al. (2013). Next we discuss identification under homogeneity of the mediator effect and introduce our new approach that relaxes this assumption. We continue by discussing cases where an appropriate experimental design can help to identify direct and indirect without having to impose additional assumptions. Thereafter, we compare the different identification strategies and show that the assumption of conditional homogeneity of the mediator effect is weaker than the sequential ignorability assumption. The second part of the paper contains the application. Here, we describe the experiment and briefly summarize the main results from Wunsch and Strobl (2018). We then discuss the implementation of the three identification strategies we compare and present our results. The last section concludes. An appendix contains the proofs of our theorems, additional information about the experiment we study in the application, and further estimation results.

## 2 Econometric framework

### 2.1 Setup

Consider the following framework which is also underlying Imai et al. (2013). We are interested in the causal effect of a binary treatment  $T$  on an outcome variable  $Y$ . There is an intermediate variable or mediator  $M$  through which the effect of the treatment on the outcome is transmitted. This paper focuses on the special case of a binary mediator because this opens up alternative possibilities for identification. Hence, we assume that the mediator  $M$  can take on the values  $m \in \{0, 1\}$ . The causal relationships that characterize our framework are illustrated in the causal diagram in Figure 1. We are interested in decomposing the treatment effect into the sum of the so-called *natural indirect effect* via the mediator of interest  $M$  (full arrows in Figure 1) and the so-called *natural direct effect* via all other possible channels (dotted arrows in Figure 1).

Figure 1: A simple causal mechanism



This paper focuses on double randomization designs that combine two randomized experiments (Pirlott and MacKinnon, 2016) and that are implemented as between-subject designs. The reason for focusing on between-subject designs is that we are interested in methods that allow studying actual behavior in the field. Within-subjects designs limit applications to either hypothetical or sequential choices. Moreover, within-subject designs require that the order of the experiments does not affect behavior in any of the two experiments (no carry-over effects; Imai et al., 2013), which can be a strong assumption. We study double randomization designs of the following form. In the baseline experiment, which we denote by  $D = 0$ , subjects are randomly assigned to the treatment of interest  $T$  and its effect on the mediator  $M$  and the outcome  $Y$  is measured. Thus, in experiment 0 the mediator is an endogenous intermediate outcome of the treatment. Experiment 0 identifies the average effect of the treatment on the mediator and the outcome. Additionally, the researcher runs an auxiliary experiment, which we denote by  $D = 1$ . In this experiment, the mediator is randomly assigned and its effect on the outcome  $Y$  is measured. Thus, experiment 1 identifies the average effect of the mediator on the outcome. We will discuss different designs that randomize different combinations of treatment and mediator values. The study subjects are randomly assigned to either experiment 0 or to experiment 1.

## 2.2 Definition of direct and indirect effects

Following Imai et al. (2013) we define direct and indirect effects within the potential outcomes framework (Rubin, 2004). The mediator can have one potential value for each combination of treatment  $T_i = t$  and experiment  $D_i = d$ ,  $M_i(t, d)$ , but only one is observed for a given subject  $i$ ,  $M_i = M_i(T_i, D_i)$ . Correspondingly, let  $Y_i(t, m, d)$  denote the potential outcome that would result if the treatment variable, the mediator and the experiment indicator equal  $t$ ,  $m$  and  $d$ , respectively. We only observe one of the potential outcomes, namely  $Y_i = Y_i(T_i, M_i(T_i, D_i), D_i)$ . By defining the observation rules in this way we are making the so-called stable unit treatment assumption of no interference between units, i.e. the potential outcomes of one unit do not depend on the values of the treatment variable, the mediator or the experiment of another unit (Cox, 1958).

We are interested in two types of effects within the baseline experiment 0. Firstly, we are interested in the overall effects of the treatment on the mediator  $M$  and the outcome of interest  $Y$ :

$$\theta_M^0 \equiv E[M_i(1, 0) - M_i(0, 0)],$$

$$\theta_Y^0 \equiv E[Y_i(1, M_i(1, 0), 0) - Y_i(0, M_i(0, 0), 0)].$$

Secondly, we are interested in disentangling the indirect effect of treatment  $T$  on the outcome  $Y$  via the mediator  $M$  from the direct effect via all other possible channels. Formally, the direct effect, which we denote by  $\delta_t^0$ , is given by

$$\delta_t^0 = E[Y_i(1, M_i(t, 0), 0) - Y_i(0, M_i(t, 0), 0)].$$

It is the expected difference in the outcome of interest due to different treatments when keeping the value of the mediator constant at  $M_i(t, 0)$ . In other words, it is the part of the effect that is not caused by the response of the mediator to the treatment. The indirect effect, which we denote by  $\gamma_t^0$ , measures the part of the overall effect of the treatment that is purely caused by the response of the mediator to the treatment while keeping everything else constant:

$$\gamma_t^0 = E[Y_i(t, M_i(1, 0), 0) - Y_i(t, M_i(0, 0), 0)].$$

It is easy to see by adding and subtracting  $E[Y_i(0, M_i(1, 0), 0)]$  that both effects add up to the total effect of the treatment on the outcome:

$$\begin{aligned} \theta_Y^0 &\equiv E[Y_i(1, M_i(1, 0), 0) - Y_i(0, M_i(0, 0), 0)] \\ &= E[Y_i(1, M_i(1, 0), 0) - Y_i(0, M_i(1, 0), 0) + Y_i(0, M_i(1, 0), 0) - Y_i(0, M_i(0, 0), 0)] \\ &= E[Y_i(1, M_i(1, 0), 0) - Y_i(0, M_i(1, 0), 0)] + E[Y_i(0, M_i(1, 0), 0) - Y_i(0, M_i(0, 0), 0)] \\ &= \delta_1^0 + \gamma_0^0. \end{aligned}$$

Instead of  $E[Y_i(0, M_i(1, 0), 0)]$ , we could also add and subtract  $E[Y_i(1, M_i(0, 0), 0)]$  yielding  $\theta_Y^0 = \delta_0^0 + \gamma_1^0$ . Following the observation rule,  $Y_i(1, M_i(1, 0), 0)$  is observed for individuals in with  $T_i = 1$  and  $D_i = 0$  while  $Y_i(0, M_i(0, 0), 0)$  is observed for individuals with  $T_i = 0$  and  $D_i = 0$ . The fundamental identification problem is that  $Y_i(0, M_i(1, 0), 0)$  (or respectively  $Y_i(1, M_i(0, 0), 0)$ ) is never observed.

### 2.3 Characterization of different double randomization designs

The designs we consider in this paper randomize subjects to two different experiments: the baseline experiment indicated by  $D_i = 0$  and the auxiliary experiment indicated by  $D_i = 1$ :



(A1) Randomization of  $D$ :

$$Y_i(t, M_i(t', d), d), M_i(t, d) \perp\!\!\!\perp D_i \quad \forall t, t', d \in \{0, 1\}.$$

The baseline experiment is the same for all designs. In this experiment we randomize the treatment  $T$  and measure its effect on the mediator  $M$  and the outcome  $Y$ :

(A2) Randomization of  $T$  in experiment 0:

$$Y_i(t, M_i(t', 0), 0), M_i(t, 0) \perp\!\!\!\perp T_i | D_i = 0 \quad \forall t, t' \in \{0, 1\}.$$

Given these two randomization assumptions the overall effects are identified from the observed mean differences in the samples of treated and non-treated subjects within experiment 0:

$$\theta_M^0 \stackrel{A1, A2}{=} E[M_i | T_i = 1, D_i = 0] - E[M_i | T_i = 0, D_i = 0], \quad (1)$$

$$\theta_Y^0 \stackrel{A1, A2}{=} E[Y_i | T_i = 1, D_i = 0] - E[Y_i | T_i = 0, D_i = 0]. \quad (2)$$

For the auxiliary experiment 1 we study four different designs. All designs have in common that they randomize the mediator:

(A3a) Randomization of  $M$  in experiment 1:

$$Y_i(t, m, 1) \perp\!\!\!\perp M_i | D_i = 1, \quad \forall t, m \in \{0, 1\}.$$

The first design we consider is the *parallel design* proposed by Imai et al. (2013). This design is characterized by randomizing subjects to one of each possible combination of treatment and mediator value in experiment 1 which means that in addition to (A3a) we have:

(A3b) Randomization of  $T$  in experiment 1:

$$Y_i(t, m, 1), M(t, 1) \perp\!\!\!\perp T_i | D_i = 1 \quad \forall t, m \in \{0, 1\}.$$

Consider the following example which corresponds to the application in Huber et al. (2017). A researcher is interested in how the labor market outcomes of unemployed workers are related to the type of caseworker they face in the public employment service and participation in a labor market program. The treatment is a less accommodating caseworker ( $T_i = 1$ ) versus a more cooperative caseworker ( $T_i = 0$ ). The mediator is participation in a labor market program ( $M_i = 1$ ) versus non-participation ( $M_i = 0$ ). In the parallel design experiment 0 would randomly assign unemployed workers to a caseworker and caseworkers would selectively assign them to a program. In experiment 1 unemployed workers would first be randomized to

caseworkers and then to the program. The parallel design results in the following observation rule for experiment 1:

$$Y_i = T_i M_i Y_i(1, 1, 1) + (1 - T_i) M_i Y_i(0, 1, 1) + T_i (1 - M_i) Y_i(1, 0, 1) + (1 - T_i) (1 - M_i) Y_i(0, 0, 1) \quad (3)$$

$$M_i = T_i M(1, 1) + (1 - T_i) M(0, 1). \quad (4)$$

It may not always be feasible to implement a parallel design, e.g. because randomizing all possible combinations of treatment and mediator with sufficiently large sample sizes may be too costly. Or there already exists an independent experiment in which the mediator has been randomized but not for (all of) the treatments considered in experiment 0. Therefore, we analyze three alternative designs. In the first of these alternative designs experiment 1 is characterized as follows. A random sample from the population of interest is drawn and each subject in this sample is randomly assigned to one of all possible values of the mediator. We call this design the *independent mediator randomization design* because the mediator is randomized independently of the treatment. For instance, in our caseworker example from above, there might exist an independently conducted experimental study where unemployed workers have been randomly assigned to the labor market program of interest. Such designs have also been discussed by Pirlott and MacKinnon (2016) but without showing formal identification results. In this design, only assumption (A3a) holds but not (A3b). The observation rule for experiment 1 is the same as in the parallel design, though.

In the second alternative design we consider, experiment 1 randomizes the mediator for one of the two treatment states of experiment 1, i.e. either  $T_i = 0$  for all  $i$  or  $T_i = 1$  for all  $i$  in experiment 0. For simplicity and without loss of generality, we will define treatment states such that  $T_i = 0$  for all  $i$  in experiment 1. For instance, in our caseworker example from above, unemployed workers might have been randomly assigned to the labor market program of interest in one regional office of the public employment service that only employs cooperative caseworkers. As the mediator is randomized conditional on treatment  $T_i = 0$  in this case, we call this design the *conditional mediator randomization design*. The randomization assumptions that characterize this experiment include (A3a) as well as (A3b), but with the restriction that (A3b) only holds for  $t = 0$ . The observation rule in experiment 1 is as follows:

$$Y_i = M_i Y_i(0, 1, 1) + (1 - M_i) Y_i(0, 0, 1) \quad (5)$$

$$M_i = M(0, 1). \quad (6)$$

The last design is a special case of the conditional mediator randomization design. It corresponds to situations where the assignment mechanism of the mediator is itself the treatment of interest, i.e. the treatment is endogenous choice of the mediator versus random assignment of the mediator. In this case,

treatment and experiment are identical, i.e. we have  $T_i = D_i$ , which is why we call this design *treatment equivalence design*. In our caseworker example, the treatment would be assignment of unemployed workers to a training program by a caseworker (experiment 0,  $T_i = D_i = 0$ ) versus random assignment to the program without the involvement of a caseworker (experiment 1,  $T_i = D_i = 1$ ). In this example a mediation framework allows investigating whether there is a value-added effect caused by the involvement of a caseworker beyond the assignment of individuals with higher expected gains, e.g. because participants are more motivated and therefore benefit more from the program because the caseworker convinced them of the value of the program. In the application in Section 4 we present another example where the assignment mechanism of the mediator is of direct interest. There we study monetary transfers to individuals who have experienced a negative income shock (outcome  $Y$ ) depending on whether risk (mediator  $M$ ) is randomly assigned and hence exogenous ( $T_i = D_i = 1$ ), or self-inflicted by making risky choices ( $T_i = D_i = 0$ ).<sup>3</sup> The randomization assumptions underlying experiment 1 in the treatment equivalence design are the same as in conditional mediator randomization design, but the observation rule is as follows:

$$Y_i = T_i M_i Y_i(1, 1, 1) + (1 - T_i) M_i Y_i(0, 1, 0) + T_i (1 - M_i) Y_i(1, 0, 1) + (1 - T_i) (1 - M_i) Y_i(0, 0, 0) \quad (7)$$

$$M_i = T_i M(1, 1) + (1 - T_i) M(0, 0). \quad (8)$$

Below we show that this design is a particularly interesting case to study because identification of the direct and indirect effects requires fewer assumptions than the other designs.

### 3 Identification

#### 3.1 Identification based on sequential ignorability

Before we discuss identification strategies that exploit double randomization we recap the strategy that is typically applied in single experiment designs which relies on the assumption of *sequential ignorability of the mediator* (Petersen et al., 2006; Flores and Flores-Lagunes, 2009; VanderWeele, 2009; Hong, 2010; Imai et al., 2010; Tchetgen Tchetgen and Shpitser, 2012; Zheng and van der Laan, 2012; Imai et al., 2013; Huber, 2014):

(A4a) Sequential ignorability of  $M$ :

$$Y_i(t', m, d) \perp\!\!\!\perp M_i | T_i = t, X_i = x \quad \forall t, t', m \in \{0, 1\}, x \in \mathcal{X}.$$

---

<sup>3</sup>Treatment equivalence designs have not been discussed in the mediation literature before. However, they share the key features of so-called doubly randomized preference trials. In these trials, subjects are randomized either to a group that can choose the treatment of interest or to a group that is randomized to the treatment. They have been used for two purposes: to assess selection bias in non-experimental settings (Rücker, 1989; Shadish et al., 2008; Joyce et al., 2017), or to assess the external validity of randomized trials (Rücker, 1989; Janevic et al., 2003; Long et al., 2008; Marcus et al., 2012).

This assumption implies that we can use subjects observed with  $M_i = m$  and  $T_i = t$  in experiment  $d$  to estimate the counterfactual outcome  $E[Y_i(t, M_i(1 - t, d), d)]$  because:

$$E[Y_i(t, M_i(1 - t, d), d)|T_i = t, M_i(1 - t, d) = m, D_i = d, X_i = x] = E[Y_i|M_i = m, T_i = t, D_i = d, X_i = x].$$

Additionally, we need to ensure that there is common support with respect to all characteristics needed for selection correction:

(A4b) Common support across treatments:

$$0 < \Pr(T_i = t|M_i = m, D_i = d, X_i = x) < 1 \quad \forall t, m \in \{0, 1\}, x \in \mathcal{X}.$$

Intuitively, we reweigh observations with  $M_i = m$  in treatment  $T_i = t$  such that they have the same distribution of covariates as those observed with  $M_i = m$  in treatment  $T_i = 1 - t$ . The common support assumption ensures that there are comparable subjects for each combination of mediator and treatment and all values of the covariates, and it is testable in the data. Under assumptions (A1), (A2), (A4a) and (A4b) the average direct effects  $\delta_t^0$  are identified for  $t \in \{0, 1\}$  and given by

$$\delta_t^0 = E \left[ \int E[Y_i|T_i = 1, M_i = m, D_i = 0, X_i = x] - E[Y_i|T_i = 0, M_i = m, D_i = 0, X_i = x] dF_{X_i|M_i=m, T_i=t, D_i=0}(x) \right]$$

in the parallel design and both mediator randomization designs. In the treatment equivalence design they are given by

$$\delta_t^0 = E \left[ \int E[Y_i|T_i = 1, M_i = m, X_i = x] - E[Y_i|T_i = 0, M_i = m, X_i = x] dF_{X_i|M_i=m, T_i=t}(x) \right].$$

For the average indirect effects we have  $\gamma_{1-t}^0 = \theta_Y^0 - \delta_t^0$  for  $t \in \{0, 1\}$  where  $\theta_Y^0 = E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$  in the treatment equivalence design and  $\theta_Y^0 = E[Y_i|T_i = 1, D_i = 0] - E[Y_i|T_i = 0, D_i = 0]$  in the other designs. Alternative identification strategies for the single-experiment case either require at least one instrumental variable (Imai et al., 2013; Yamamoto, 2014; Park, 2015; Frölich and Huber, 2017) or rely on common trend assumptions imposed on certain subpopulations that require multiple observations over time (Deuchert et al., 2016) and may only identify the effects of interest for certain sub-populations.

## 3.2 Exploiting double randomization

### 3.2.1 Parallel design

Imai et al. (2013) have proposed an identification strategy for the parallel design that we use as a benchmark for the discussion of alternative designs and identification strategies. In the following, we recap this strategy before we discuss alternatives. In the parallel design, experiment 1 randomizes each

subject to one of all possible combinations of treatment and mediator value. In the binary mediator and treatment case this implies that  $E[Y_i(t, m, 0)]$  is identified and equal to  $E[Y_i | T_i = t, M_i = m, D_i = 0]$  for all combinations of  $t, m \in \{0, 1\}$ . Imai et al. (2013) show that identification of the direct and indirect effects can be achieved based on the randomization assumptions that characterize this design, i.e. (A1), (A2), (A3a) and (3b), together with the following two assumptions:

(A5) Consistency:

$$Y_i(t, M_i(t, 0), 0) = Y_i(t, m, 1) \quad \text{if } M_i(t, 0) = m \quad \forall t, m \in \{0, 1\}.$$

The consistency assumption allows us to write  $Y_i(t, m, d)$  as  $Y_i(t, m)$  for all  $t, m, d \in \{0, 1\}$ . The consistency assumption requires that individual potential outcomes for a given value of the mediator and the treatment are the same in both experiments, i.e. that experiment 1 has no direct effect on the potential outcomes. In other words, it does not matter how the potential outcomes come about (from experiment 0 or 1) as long as the treatment and mediator values are the same. Imai et al. (2013) emphasize that this can be a strong assumption. To see this consider our caseworker example from above where a researcher compares the labor market outcomes of unemployed workers assigned to a less accommodating caseworker ( $T_i = 1$ ) versus a more cooperative caseworker ( $T_i = 0$ ) with participation in a labor market program ( $M_i = 1$ ) versus non-participation ( $M_i = 0$ ) as the mediator. The consistency assumption requires that subjects do not behave differently when program assignment is random rather than the caseworker's decision. This excludes demotivation effects when being denied access to the program without a reason under randomization or motivation effects that lead to higher effort in the program under caseworker assignment. The second assumption is

(A6) No causal interaction between treatment and mediator:

$$Y_i(t, m, d) - Y_i(t, m', d) = Y_i(t', m, d) - Y_i(t', m', d) \quad \forall t \neq t' \text{ and } t, t', m, m', d \in \{0, 1\}.$$

Assumption (A6) directly implies that  $\gamma_0^1 = \gamma_1^1$  and  $\delta_0^1 = \delta_1^1$ . How strong this assumption is will depend on the application. In our caseworker example program participation may have a smaller effect for the same individual if assigned by a less accommodating caseworker compared to assignment by a more cooperative caseworker for the following reason. Assume that the worker is initially opposed to participation. The cooperative worker will try to convince the worker of the value of the program. If successful, the worker may be more motivated and pay more attention to the program which may result in higher gains from the program compared to the case where the same worker is forced into the program against his will by a non-cooperative caseworker. In this case the assumption of no interaction effect between treatment and mediator would be violated.

Under assumptions (A1), (A2), (A3a), (A3b), (A5) and (A6) the average direct and indirect effects  $\delta_t^0$  and  $\gamma_t^0$  are identified for  $t \in \{0, 1\}$  and given by:

$$\delta_t^0 = E[Y_i|T_i = 1, M_i = m, D_i = 1] - E[Y_i|T_i = 0, M_i = m, D_i = 1] \quad (9)$$

$$\gamma_t^0 = \theta_Y^0 - \delta_{1-t}^0 \quad (10)$$

for  $m \in \{0, 1\}$  where  $\theta_Y^0 = E[Y_i|T_i = 1, D_i = 0] - E[Y_i|T_i = 0, D_i = 0]$ . The direct and indirect effects do not depend on the treatment, i.e.  $\delta_0^0 = \delta_1^0$  and  $\gamma_0^0 = \gamma_1^0$ , and they should be the same for different values of the mediator  $m$ . This is a direct consequence of assumption (A6) of no causal interaction between treatment and mediator. From the estimand of the direct effect it is easy to see then that assumptions (A1), (A2), (A3a), (A3b) and (A6) together imply that the following testable condition must be satisfied in the data:

$$\begin{aligned} (C1): \quad & E[Y_i|T_i = 1, M_i = m, D_i = 1] - E[Y_i|T_i = 0, M_i = m, D_i = 1] \\ &= E[Y_i|T_i = 1, M_i = 1 - m, D_i = 1] - E[Y_i|T_i = 0, M_i = 1 - m, D_i = 1] \\ &\Leftrightarrow E[Y_i|T_i = t, M_i = 1, D_i = 1] - E[Y_i|T_i = t, M_i = 0, D_i = 1] \\ &= E[Y_i|T_i = 1 - t, M_i = 1, D_i = 1] - E[Y_i|T_i = 1 - t, M_i = 0, D_i = 1] \end{aligned}$$

for  $t, m \in \{0, 1\}$  (see also Imai et al. 2013). The average treatment (mediator) effect conditional on a particular value of the mediator (treatment) in experiment 1 must be the same for all values of the mediator (treatment). The testability of this condition is an attractive feature of the parallel design with a discrete mediator. It is a direct implication of assumption (A6) of no causal interaction between treatment and mediator. Hence, condition (C1) provides a falsification test for validity of the key assumption (A6) of the parallel design when the mediator is a binary (or more generally a discrete) variable. Another implication of this finding is that identification only requires randomization of the treatment for one value of the mediator in experiment 1. Randomization for more than one mediator value is useful for assessing the plausibility of assumption (A6) but it is not required for identification.

### 3.2.2 Designs that randomize the mediator

In designs that do not randomize the treatment for at least one mediator value identification of the direct and indirect effects is more difficult (Pirlott and MacKinnon, 2016). Moreover, the key assumption of no causal interaction between treatment and mediator (A6) might be violated in parallel designs as condition (C1) might be rejected by the data. In the following we discuss alternative identification strategies that can be applied when the auxiliary experiment randomizes the mediator.

To understand the nature of the identification problem it is useful to think about the different possible responses to the treatment in the baseline experiment 0. Some subjects will have the same value of the

mediator independent of treatment status. Borrowing from the terminology of Angrist et al. (1996) we distinguish between so-called always takers ( $\tau_i = (1, 1)$ ) with  $M_i(0, 0) = M_i(1, 0) = 1$  and never takers ( $\tau_i = (0, 0)$ ) with  $M_i(0, 0) = M_i(1, 0) = 0$ . We arbitrarily label subjects as compliers ( $\tau_i = (0, 1)$ ) if they switch from  $M_i(0, 0) = 0$  to  $M_i(1, 0) = 1$ . Correspondingly, subjects who switch from  $M_i(0, 0) = 1$  to  $M_i(1, 0) = 0$  are labeled as defiers ( $\tau_i = (1, 0)$ ). The corresponding potential outcomes and indirect effects are summarized in Table 1.

Table 1: Potential responses to the treatment in experiment 0

Type	$\tau_i$	$M_i(0, 0)$	$M_i(1, 0)$	$Y_i(t, M_i(1, 0), 0)$	$Y_i(t, M_i(0, 0), 0)$	$\gamma_{t,i}^0$
Always taker	(1, 1)	1	1	$Y_i(t, 1, 0)$	$Y_i(t, 1, 0)$	0
Never taker	(0, 0)	0	0	$Y_i(t, 0, 0)$	$Y_i(t, 0, 0)$	0
Complier	(0, 1)	0	1	$Y_i(t, 1, 0)$	$Y_i(t, 0, 0)$	$Y_i(t, 1, 0) - Y_i(t, 0, 0)$
Defier	(1, 0)	1	0	$Y_i(t, 0, 0)$	$Y_i(t, 1, 0)$	$Y_i(t, 0, 0) - Y_i(t, 1, 0)$

Always-takers and never-takers do not change the value of the mediator in response to the treatment. Therefore, the indirect effect  $\gamma_{t,i}^0$  is zero for these subjects. As a consequence, the indirect effect equals

$$\begin{aligned} \gamma_t^0 &= E[Y_i(t, 1, 0) - Y_i(t, 0, 0) | \tau_i = (0, 1)] Pr(\tau_i = (0, 1)) \\ &\quad - E[Y_i(t, 1, 0) - Y_i(t, 0, 0) | \tau_i = (1, 0)] Pr(\tau_i = (1, 0)). \end{aligned} \quad (11)$$

Identification of the indirect effect requires identification of the mediator effect conditional on type, i.e. of  $E[Y_i(t, 1, 0) - Y_i(t, 0, 0) | \tau_i = \tau]$ . In a between-subject design we only observe each subject in one of the two treatment states. Hence, we cannot determine which subject belongs to which type. As a consequence, we cannot identify the indirect effect without further assumptions. Note, however, that experiment 1 creates exogenous variation in the mediator, which should be useful for identification of the mediator effect. One prerequisite for this to work is that experiment 1 is informative for the potential outcomes we need to identify for experiment 0. Therefore, we need to impose the consistency assumption (A5) in all double randomization designs with one exception. In the treatment equivalence design, where  $T_i = D_i$ , potential outcomes only depend on  $t$  and  $m$  by construction rather than by assumption (A5). As a consequence, the indirect effect simplifies for all designs to

$$\begin{aligned} \gamma_t &= E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = (0, 1)] Pr(\tau_i = (0, 1)) \\ &\quad - E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = (1, 0)] Pr(\tau_i = (1, 0)). \end{aligned} \quad (12)$$

Now note that experiment 1 identifies the average mediator effect on the outcome  $Y_i$ . It is easy to see then, that experiment 1 is useful for identification under the following effect homogeneity assumption:

(A7) Homogeneous effect of  $M$  on  $Y$ :

$$E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = \tau] = E[Y_i(t, 1) - Y_i(t, 0)] \quad \forall t \text{ and } \tau \in \{(0, 1), (1, 0)\}.$$

Under assumption (A7) the indirect effect simplifies to the product of the average mediator effect and the average effect of the treatment on the mediator:

$$\begin{aligned}
\gamma_t &= E[Y_i(t, 1) - Y_i(t, 0)][Pr(\tau_i = (0, 1)) - Pr(\tau_i = (1, 0))] \\
&= E[Y_i(t, 1) - Y_i(t, 0)][Pr(\tau_i = (0, 1)) + Pr(\tau_i = (1, 1)) - Pr(\tau_i = (1, 0) - Pr(\tau_i = (1, 1)))] \\
&= E[Y_i(t, 1) - Y_i(t, 0)]E[M_i(1, 0) - M_i(0, 0)].
\end{aligned}$$

The average effect of the treatment on the mediator  $\theta_M^0 = E[M_i(1, 0) - M_i(0, 0)]$  is directly identified from experiment 0 based on assumptions (A1) and (A2). Identification of the mediator effect  $E[Y_i(t, 1) - Y_i(t, 0)]$  differs by design. In the parallel design, the mediator effect is identified under assumptions (A1), (A3a), (A3b) and (A5) for both treatments, i.e.

$$E[Y_i(t, 1) - Y_i(t, 0)] = E[Y_i|T_i = t, M_i = 1, D_i = 1] - E[Y_i|T_i = t, M_i = 0, D_i = 1]$$

for  $t \in \{0, 1\}$ . In the independent mediator randomization design  $E[Y_i(t, 1) - Y_i(t, 0)]$  is not identified directly. Here, experiment 1 identifies

$$E[Y_i(t, 1) - Y_i(t, 0)|T_i = t] = E[Y_i|T_i = t, M_i = 1, D_i = 1] - E[Y_i|T_i = t, M_i = 0, D_i = 1].$$

under assumptions (A1), (A3a) and (A5). For identification of average mediator effect we additionally need to impose the assumption of no causal interaction between treatment and mediator (A6), which implies that  $E[Y_i(t, 1) - Y_i(t, 0)] = E[Y_i(1 - t, 1) - Y_i(1 - t, 0)]$  for all  $t \in \{0, 1\}$ . In this case we have

$$\begin{aligned}
E[Y_i(t, 1) - Y_i(t, 0)] &= \sum_{t \in \{0, 1\}} E[Y_i(t, 1) - Y_i(t, 0)|T_i = t]Pr(T_i = t) \\
&= \sum_{t \in \{0, 1\}} [E[Y_i|T_i = t, M_i = 1, D_i = 1] - E[Y_i|T_i = t, M_i = 0, D_i = 1]]Pr(T_i = t) \\
&= E[Y_i|M_i = 1, D_i = 1] - E[Y_i|M_i = 0, D_i = 1]
\end{aligned}$$

for  $t \in \{0, 1\}$ . In the conditional mediator randomization design,  $E[Y_i(t, 1) - Y_i(t, 0)]$  is directly identified from experiment 1 under assumptions (A1), (A3a), (A3b) and (A5) for  $t = 0$ . In the treatment equivalence design,  $E[Y_i(t, 1) - Y_i(t, 0)]$  is directly identified from experiment 1 under assumptions (A1), (A3a) and (A3b) for  $t = 1$ . The following theorem summarizes the identification results for all designs:

**Theorem 1:** Under assumptions (A1), (A2), (A3a), (A3b), (A5) and (A7) the following average indirect effects are identified in the parallel design and given by:

$$\gamma_t = \{E[Y_i|T_i = t, M_i = 1, D_i = 1] - E[Y_i|T_i = t, M_i = 0, D_i = 1]\}\theta_M^0 \quad (13)$$



for  $t \in \{0, 1\}$  where  $\theta_M^0 = E[M_i|T_i = 1, D_i = 0] - E[M_i|T_i = 0, D_i = 0]$ . In the mediator randomization designs the following average indirect effects are identified and given by:

$$\gamma_t = \{E[Y_i|M_i = 1, D_i = 1] - E[Y_i|M_i = 0, D_i = 1]\}\theta_M^0 \quad (14)$$

where  $\theta_M^0 = E[M_i|T_i = 1, D_i = 0] - E[M_i|T_i = 0, D_i = 0]$  under assumptions (A1), (A2), (A3a), (A3b), (A5), (A6), (A7) and for  $t \in \{0, 1\}$  in the independent mediator randomization design, and under assumptions (A1), (A2), (A3a), (3b), (A5), (A7) and for  $t = 0$  in the conditional mediator randomization design. In the treatment equivalence design the following average indirect effect is identified and given by:

$$\gamma_1 = \{E[Y_i|M_i = 1, T_i = 1] - E[Y_i|M_i = 0, T_i = 1]\}\{E[M_i|T_i = 1] - E[M_i|T_i = 0]\} \quad (15)$$

under assumptions (A1), (A2), (A3a), (A3b) and (A7). For the average direct effects we have  $\delta_{1-t} = \theta_Y^0 - \gamma_t$  whenever  $\gamma_t$  is identified where  $\theta_Y^0 = E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$  in the treatment equivalence design and  $\theta_Y^0 = E[Y_i|T_i = 1, D_i = 0] - E[Y_i|T_i = 0, D_i = 0]$  in the other designs.

The proof is given in Appendix A.1. Note that in the parallel design, we have replaced the assumption of no causal interaction between treatment and mediator (A6) by the effect homogeneity assumption (A7). This also implies that we no longer impose  $\gamma_1 = \gamma_0$  and  $\delta_0 = \delta_1$ . Moreover, because the mediator is randomized under both treatments in experiment 1, all direct and indirect effects of interest are identified, which shows the identifying power of this design. The independent mediator randomization design requires very strong assumptions as both the consistency assumption (A5) and the assumption of no causal interaction between treatment and mediator (A6) are needed in addition to the effect homogeneity assumption (A7) and we have  $\gamma_1 = \gamma_0$  and  $\delta_0 = \delta_1$  as a result. As in the parallel design, the conditional mediator randomization design requires the consistency assumption (A5) for  $t = 0$  but not the assumption of no causal interaction between treatment and mediator (A6) unless one wishes to identify  $\gamma_1$  and  $\delta_0$  as well. The treatment equivalence design neither requires the consistency assumption (A5) nor the assumption of no causal interaction between treatment and mediator (A6). As Imai et al. (2013) emphasize, both can be strong assumptions which makes this design attractive and an interesting case to study. However,  $\gamma_0$  and  $\delta_1$  are not identified unless one is willing to additionally impose assumption (A6) of no interaction effects.

A direct implication of Theorem 1 and its proof is that any double randomization design that identifies the average effect of the treatment on the mediator  $E[M_i(1, 0) - M_i(0, 0)]$  and the average mediator effect  $E[Y_i(t, 1) - Y_i(t, 0)]$  opens up the possibility for identification of the direct and indirect effect for at least one treatment state based on the assumption of a homogeneous mediator effect (A7). In the mediator randomization designs and the treatment equivalence design, identification based on just the consistency

assumption (A5) and the assumption of no causal interaction between treatment and mediator (A6) does not suffice for identification because the treatment is not randomized for a given mediator value. Therefore, identification based on a homogeneous mediator effect is the only alternative to sequential ignorability in these designs. Yet, homogeneity of the mediator effect is a strong assumption that is likely to be violated in many applications. At least, we can easily check whether homogeneity of the effect of the mediator on the outcome can be rejected in the data for observed (but not unobserved) characteristics by estimating the effect in subsamples defined by certain subject characteristics. I.e. we can test whether the following condition holds in the data:

$$(C2): \quad E[Y_i|T_i = t, M_i = 1, D_i = 0] - E[Y_i|T_i = t, M_i = 0, D_i = 0] \\ = E[Y_i|T_i = t, M_i = 1, D_i = 0, X_i = x] - E[Y_i|T_i = t, M_i = 0, D_i = 0, X_i = x]$$

for at least some  $x \in \mathcal{X}$  and the values of  $t$  for which the mediator has been randomized in experiment 1, which varies across designs. Now what if the data reject condition (C2) for some covariates? Shall we give up on studying causal mechanisms in this case? Or do we have to turn back to identification based on sequential ignorability (A4a,b) without exploiting the random variation in the mediator generated in experiment 1? Remember that

$$\gamma_t = E[Y_i(t, 1) - Y_i(t, 0)|\tau_i = (0, 1)]Pr(\tau_i = (0, 1)) - E[Y_i(t, 1) - Y_i(t, 0)|\tau_i = (1, 0)]Pr(\tau_i = (1, 0)).$$

The identification problem for the indirect effects originates from the inability to observe types, i.e. we do not know who is an always-taker, never-taker, complier or defier. As a result, we cannot calculate the mediator effects in experiment 1 conditional on type, which directly implies that we are unable to capture effect heterogeneity with respect to type. However, intuitively it should be possible to handle heterogeneity with respect to all or at least a subset of observed characteristics  $X_i \in \mathcal{X}$ . Consider the following assumption of conditional homogeneity of the mediator effect:

(A8) Homogeneous effect of  $M$  on  $Y$  conditional on  $X$ :

$$E[Y_i(t, 1) - Y_i(t, 0)|\tau_i = \tau, X_i = x] = E[Y_i(t, 1) - Y_i(t, 0)|X_i = x] \quad \forall t \text{ and } \tau \in \{(0, 1), (1, 0)\}, x \in \mathcal{X}.$$

The difference to the effect homogeneity assumption (A7) is that we allow for arbitrary effect heterogeneity with respect to the characteristics  $X$ . Angrist and Fernandez-Val (2013) and Aronow and Carnegie (2013) make a similar assumption to recover the average treatment effect (and other parameters) from the local average treatment effect (LATE) that is identified with an instrumental variable (IV) approach in the presence of heterogeneous treatment effects. In fact, they consider the reverse problem of the one we face in the mediation case. The LATE is the effect for the compliers (with defiers assumed away) from which

they seek to identify the average treatment effect for the population based on conditional LATE's. In our case, the auxiliary experiment identifies the average effect of the mediator and we seek to identify the mediator effect for the compliers and defiers, where defiers do not have to be assumed away. This difference is important because we know that the compliers and defiers are a subpopulation of the population we use to estimate the average mediator effect due to randomization of the mediator in experiment 1. In contrast, if subjects with certain characteristics are never compliers in the LATE framework, then it is impossible to reconstruct the average treatment effect from conditional LATE's because we can only construct the latter for the covariates conditional on which the monotonicity assumption holds.

Applying the law of iterated expectations and Bayes' theorem it can be shown that under assumption (A8) the indirect effect becomes

$$\begin{aligned}\gamma_t &= \int E[Y_i(t, 1) - Y_i(t, 0)|X_i = x][Pr(\tau_i = (0, 1)|X_i = x) - Pr(\tau_i = (1, 0)|X_i = x)]f_{X_i}(x)dx \\ &= \int E[Y_i(t, 1) - Y_i(t, 0)|X_i = x]E[M_i(1, 0) - M_i(0, 0)|X_i = x]f_{X_i}(x)dx,\end{aligned}$$

which is the conditional on  $X$  equivalent of what we obtain under a globally homogeneous mediator effect. As a result, we need to compute the mediator effect conditional on  $X$ , which is identified from experiment 1 due to randomization of the mediator, as well as the effect of the treatment on the mediator conditional on  $X$ , which is identified from experiment 0 due to randomization of the treatment, and then aggregate over the distribution of  $X$ . In the following, we show that identification of the same direct and indirect effects as in Theorem 1 can be achieved by replacing the assumption of a homogeneous mediator effect (A7) with the assumption of conditional effect homogeneity (A8).

**Theorem 2:** Let  $p_{M,i}^d(t, x) \equiv Pr(M_i = 1|T_i = t, D_i = d, X_i = x)$  and  $p_{M,i}^d(x) \equiv Pr(M_i = 1|D_i = d, X_i = x)$  for all  $t, d \in \{0, 1\}, x \in \mathcal{X}$ . Under assumptions (A1), (A2), (A3a), (A3b), (A5) and (A8) the following average direct and indirect effects are identified in the parallel design and given by:

$$\gamma_t = E \left[ \left( \frac{Y_i M_i}{p_{M,i}^1(t, x)} - \frac{Y_i(1 - M_i)}{1 - p_{M,i}^1(t, x)} \right) (p_{M,i}^0(1, x) - p_{M,i}^0(0, x)) | T_i = t \right] \quad (16)$$

for  $t \in \{0, 1\}$ . In the mediator randomization designs the following average direct and indirect effects are identified and given by:

$$\gamma_t = E \left[ \left( \frac{Y_i M_i}{p_{M,i}^1(x)} - \frac{Y_i(1 - M_i)}{1 - p_{M,i}^1(x)} \right) (p_{M,i}^0(1, x) - p_{M,i}^0(0, x)) | T_i = t \right] \quad (17)$$

under assumptions (A1), (A2), (A3a), (A3b), (A5), (A6) and (A8) and for  $t \in \{0, 1\}$  in the independent mediator randomization design, and under assumptions (A1), (A2), (A3a), (3b), (A5) and (A8) and for  $t = 0$  in the conditional mediator randomization design. In the treatment equivalence design the following

average direct and indirect effects are identified and given by:

$$\gamma_1 = E \left[ \left( \frac{Y_i M_i}{p_{M,i}^1(x)} - \frac{Y_i(1 - M_i)}{1 - p_{M,i}^1(x)} \right) (p_{M,i}^1(x) - p_{M,i}^0(x)) | T_i = 1 \right] \quad (18)$$

under assumptions (A1), (A2), (A3a), (A3b) and (A8). For the average direct effects we have  $\delta_{1-t} = \theta_Y^0 - \gamma_t$  where  $\theta_Y^0 = E[Y_i | T_i = 1] - E[Y_i | T_i = 0]$  in the treatment equivalence design and  $\theta_Y^0 = E[Y_i | T_i = 1, D_i = 0] - E[Y_i | T_i = 0, D_i = 0]$  in the other designs.

The proof is given in Appendix A.2. Note that due to the randomization of the mediator in experiment 1, we have  $p_{M,i}^1(t, x) = p_{M,i}^1(t) \equiv \Pr(M_i = 1 | T_i = t, D_i = 1)$  for the parallel design and  $p_{M,i}^1(x) = p_{M,i}^1 \equiv \Pr(M_i = 1 | D_i = 1)$  for the other designs, i.e. we do not need to condition on  $X$ .<sup>4</sup> However, we will keep the possible dependence on  $X$  explicit because the estimator we derive on the basis of Theorem 2 will then directly allow correcting for small sample covariate imbalances, which can be relevant in applications. The proof of Theorem 2 reveals that any double randomization design that identifies the average effect of the treatment on the mediator  $E[M_i(1, 0) - M_i(0, 0)]$  and the average mediator effect  $E[Y_i(t, 1) - Y_i(t, 0)]$  opens up the possibility of identification of the direct and indirect effect for at least one treatment state based on the assumption of conditional homogeneity of the mediator effect (A8).

We propose a simple semi-parametric propensity score weighting estimator for the estimands of the indirect effects derived in Theorem 2. Specifically, we use normalized versions of the sample analogs of the components of the estimands, such that the weights of the observations add up to unity, as advocated in Imbens (2004) and Busso et al. (2014). The estimator for the parallel design is given by

$$\hat{\gamma}_t = \frac{\sum_{i:T_i=t, D_i=1} w_i^1(t, x) Y_i M_i}{\sum_{i:T_i=t, D_i=1} w_i^1(t, x) M_i} - \frac{\sum_{i:T_i=t, D_i=1} w_i^0(t, x) Y_i (1 - M_i)}{\sum_{i:T_i=t, D_i=1} w_i^0(t, x) (1 - M_i)} \quad (19)$$

$$w_i^0(t, x) = \frac{\hat{p}_{M,i}^0(1, x) - \hat{p}_{M,i}^0(0, x)}{1 - \hat{p}_{M,i}^1(t, x)}, \quad w_i^1(t, x) = \frac{\hat{p}_{M,i}^0(1, x) - \hat{p}_{M,i}^0(0, x)}{\hat{p}_{M,i}^1(t, x)} \quad (20)$$

for  $t \in \{0, 1\}$  where  $\hat{p}_{M,i}^d(\cdot)$  denotes the estimate of  $p_{M,i}^d(\cdot)$ . These estimates are also called propensity scores. For the mediator randomization designs and the treatment equivalence design we have

$$\hat{\gamma}_t = \frac{\sum_{i:D_i=1} w_i^1(x) Y_i M_i}{\sum_{i:D_i=1} w_i^1(x) M_i} - \frac{\sum_{i:D_i=1} w_i^0(x) Y_i (1 - M_i)}{\sum_{i:D_i=1} w_i^0(x) (1 - M_i)} \quad (21)$$

for  $t \in \{0, 1\}$  in the independent mediator randomization design,  $t = 0$  in the conditional mediator ran-

---

<sup>4</sup>Also note that this means that we do not need to impose common support for these probabilities. Moreover, lack of support with respect to  $p_{M,i}^0(\cdot)$  is unproblematic as these probabilities do not enter the denominators.

domization design and  $t = 1$  in the treatment equivalence design where

$$w_i^0(x) = \frac{\hat{p}_{M,i}^0(1,x) - \hat{p}_{M,i}^0(0,x)}{1 - \hat{p}_{M,i}^1(x)}, \quad w_i^1(x) = \frac{\hat{p}_{M,i}^0(1,x) - \hat{p}_{M,i}^0(0,x)}{\hat{p}_{M,i}^1(x)} \quad (22)$$

in the mediator randomization designs and

$$w_i^0(x) = \frac{\hat{p}_{M,i}^1(x) - \hat{p}_{M,i}^0(x)}{1 - \hat{p}_{M,i}^1(x)}, \quad w_i^1(x) = \frac{\hat{p}_{M,i}^1(x) - \hat{p}_{M,i}^0(x)}{\hat{p}_{M,i}^1(x)} \quad (23)$$

in the treatment equivalence design. The propensity scores,  $\hat{p}_{M,i}^d(t,x)$  and  $p_{M,i}^d(x)$ , are consistent parametric estimates of, respectively,  $Pr(M_i = 1|T_i = t, D_i = d, X_i = x)$  and  $Pr(M_i = 1|D_i = d, X_i = x)$  in the sample with  $T_i = t, D_i = d$  (e.g. using a probit model). The final estimate only uses subjects from experiment 1. To obtain  $\hat{p}_{M,i}^0(t,x)$  for the subjects in experiment 1, we use the coefficients from the propensity score models estimated based on the subjects from experiment 0 and predict the respective probabilities using the covariates of the subjects in experiment 1.<sup>5</sup>

Our semi-parametric estimators (into which the propensity scores enter parametrically) can be expressed as sequential GMM estimators where propensity score estimation represents the first step and effect estimation the second step. Newey (1984) developed a procedure to derive the asymptotic distribution of estimators of this type. It requires imposing a set of standard regularity conditions that are considered to be weak (Newey, 1984). It follows from his results that the proposed estimators are  $\sqrt{N}$ -consistent.<sup>6</sup> Furthermore, the weighting estimators are sufficiently smooth for the bootstrap to be consistent for inference (see also Angrist and Fernandez-Val (2013)). Therefore, we estimate the standard errors in the application in Section 4 by 1999 bootstrap replications.

### 3.2.3 Treatment equivalence design as a special case

To complete the discussion of possible ways to identify direct and indirect effects, we show in the following that appropriately implemented treatment equivalence designs allow for identification without having to impose additional assumptions. In treatment equivalence designs, the treatment is endogenous choice ( $T = 0$ ) versus randomization ( $T = 1$ ) of the mediator. In principle, there are different ways to implement the randomization of the mediator. The most intuitive one would be to randomly assign subjects to one or the other treatment at the beginning of the experiment. Now consider the following alternative. At the beginning of the experiment, all subjects are randomly assigned a treatment status  $T_i \in \{0, 1\}$ . As the next step, all subjects are asked to state their preferred value of the mediator. The subjects with

<sup>5</sup>In a recent paper, Hsu et al. (2017) propose non-parametric weighting estimators for natural direct and indirect effects based on the sequential ignorability assumption. This method does not extend to our case because we need to make this out-of-sample prediction. This is also the case for the estimator derived by Deuchert and Wunsch (2014), which has the same structure as our estimator.

<sup>6</sup>Deuchert and Wunsch (2014) derive an estimator that has the same structure as our estimator although it has been derived from different assumptions and in a different context. They show  $\sqrt{N}$ -consistency of this estimator in the appendix.

treatment status  $T_i = 0$  receive their preferred option, while for subjects with treatment status  $T_i = 1$  a randomization device determines whether they receive their preferred option or the alternative.<sup>7</sup> In this case, we observe the mediator value under choice,  $M_i(0)$ , for all subjects in the experiment including those with  $T_i = 1$ . As a result, we can identify

$$E[Y_i(1, M_i(0))] = E[Y_i(1, 0)|M_i(0) = 0]Pr(M_i(0) = 0) + E[Y_i(1, 1)|M_i(0) = 1]Pr(M_i(0) = 1)$$

using average observed outcomes of subjects for whom randomly assigned and preferred mediator value coincide:

$$E[Y_i(1, M_i(0))] = E[Y_i|T_i = 1, M_i = M_i(0) = 0]Pr(M_i(0) = 0) + E[Y_i|T_i = 1, M_i = M_i(0) = 1]Pr(M_i(0) = 1).$$

This allows identifying  $\delta_0$  and  $\gamma_1$  without having to impose additional assumptions. Eliciting  $M_i(0)$  for subjects with  $T_i = 1$  this way may not always be possible due to the nature of mediator. An alternative would be to use the intuitive design but ask subjects with randomized mediator value whether this is the value they would have chosen for themselves. In this case, the researcher needs to ensure that ex post stated preferences do not differ systematically from the actual choices made in the actual choice treatment (see Wunsch and Strobl, 2018, for an example).

### 3.3 Comparing alternative identification strategies

The identification strategies discussed above imply that for each of the above double randomization designs there are at least three possible identification strategies: (i) identification based on the sequential ignorability assumption (approach 1), (ii) identification that exploits the double randomization based on the assumption of a homogeneous mediator effect (approach 2), and (iii) the approach that requires a homogeneous mediator effect conditional on observed characteristics (approach 3). For the parallel design, there additionally exists the approach of Imai et al. (2013) with the key identifying assumption of no interaction of treatment and mediator (approach 4). Moreover, the treatment equivalence design may not require any additional assumptions if  $M_i(0)$  is measured for all subjects (approach 5). Table 2 summarizes the required identifying assumptions by approach and design. When we focus on the three approaches that work in all designs, we can rank the four designs in terms of the identifying assumptions they require. Approach 1 based on sequential ignorability (A4a/b) is equally demanding for all designs. Hence, we focus on the other approaches when ranking the designs. The treatment equivalence design is least demanding because it requires neither the consistency assumption (A5) nor the assumption of no causal interaction between treatment and mediator (A6), which makes it an interesting case to study.

---

<sup>7</sup>See Karlan and Zinman (2009) and Dal et al. (2010) for examples of such designs, which have not been used in the context of mediation analysis, though.

The parallel design and the conditional mediator design require the former but not the latter, while the independent mediator randomization design requires both, which makes it the most demanding design.

Table 2: Identifying assumptions by approach and design

Approach	Assumption	Parallel design	Mediator randomization design	Treatment equivalence design
Approach 1	<b>(A4a/b) Sequential ignorability*</b>	x	x	x
	(A1) Randomization of $D$ and of $T$ for $D = 1$	x	x	x
Approach 2	<b>(A7) Homogeneous mediator effect</b>	x	x	x
	(A1) Randomization of $D$	x	x	x
	(A2) Randomization of $T$ for $D = 0$	x	x	x
	(A3a/b) Randomization of $M$ for $D = 1$	x	x	x
	(A5) Consistency	x	x	x
	(A6) No interaction between $T$ and $M$		x	
Approach 3	<b>(A8) Homogeneous mediator effect conditional on <math>X^*</math></b>	x	x	x
	(A1) Randomization of $D$	x	x	x
	(A2) Randomization of $T$ for $D = 0$	x	x	x
	(A3a/b) Randomization of $M$ for $D = 1$	x	x	x
	(A5) Consistency	x	x	x
	(A6) No interaction between $T$ and $M$		x	
Approach 4	<b>(A6) No interaction between <math>T</math> and <math>M</math></b>	x		
	(A1) Randomization of $D$	x		
	(A2) Randomization of $T$ for $D = 0$	x		
	(A3a/b) Randomization of $M$ for $D = 1$	x		
	(A5) Consistency	x		
Approach 5	<b>Direct measurement of <math>M_i(0)</math> for subjects in <math>T = 1</math></b>			x
	(A1) Randomization of $D$			x
	(A2) Randomization of $T$ for $D = 0$			x
	(A3a/b) Randomization of $M$ for $D = 1$			x

Note: The key identifying assumption for each approach is indicated in bold letters. \*Plus common support.

A direct consequence of the availability of multiple identification strategies is that this allows mutually testing them against each other because they will only yield the same results if all of the identifying assumptions underlying the tested approaches hold. Moreover, if we have reason to believe that one of these strategies is plausible, then we can test the validity of the other strategies. For example, if in the parallel design identification based on consistency and no interaction of treatment and mediator is plausible, then we can test whether identification based on sequential ignorability yields the same results, which would imply that validity of approach 1 cannot be rejected in the data. Such tests are useful because it may not always be possible to implement double randomization experiments on a sufficiently large scale for each question of interest, e.g. because it would be prohibitively costly. In particular, knowing which covariates are required or sufficient to justify identification based on the sequential ignorability assumption from a smaller scale double randomization experiment is extremely useful to be able to identify causal mechanisms in cases where the mediator is not randomly assigned. In the following we discuss whether it is possible to rank the different identification strategies.

It is obvious from Table 2 and the discussion in the last section that approach 3 based on conditional homogeneity of the mediator effect (A8) makes weaker assumptions than approach 2 based on global homogeneity of the mediator effect (A7). In the parallel design, we can replace these effect homogeneity assumptions with the assumption of no causal interaction between treatment and mediator (A6) but it is unclear which set of identifying assumptions is more restrictive in a given application. The same holds for identification based on sequential ignorability (A4a/b) in the parallel design, which cannot be ranked against identification based on no interaction effects (A6). Also, it is unclear whether the sequential

ignorability assumption (A4a) or the assumption of global homogeneity of the mediator effect (A7) is more restrictive, which implies that we cannot rank approaches 1 and 2 in all designs. However, in the following we show that the assumption of conditional homogeneity of the mediator effect (A8) is weaker than the sequential ignorability assumption (A4a) which allows us to rank approaches 1 and 3 under certain conditions.

In the treatment equivalence design and under the consistency assumption (A5) in the other designs the indirect effect is given by

$$\begin{aligned}\gamma_t &= E[Y_i(t, M(1)) - Y_i(t, M(0))] \\ &= E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = (0, 1)]Pr(\tau_i = (0, 1)) - E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = (1, 0)]Pr(\tau_i = (1, 0)).\end{aligned}$$

Let  $X$  denote the vector of covariates that ensures that the sequential ignorability assumption (A4a) holds. Now note that type  $\tau_i$  is determined by selection of a mediator value  $m$  depending on treatment  $t$ . This implies that if the vector  $X$  includes all relevant confounders, then

$$\begin{aligned}E[Y_i(t, m) | \tau_i = \tau] &= \int E[Y_i(t, m) | X_i = x, \tau_i = \tau] f_{X_i | \tau_i = \tau}(x) \\ &= \int E[Y_i(t, m) | X_i = x] f_{X_i | \tau_i = \tau}(x).\end{aligned}$$

This implies, firstly, that the assumption of conditional homogeneity of the mediator effect (A8) holds automatically if the sequential ignorability assumption (A4a) holds. Secondly, it implies that for validity of assumption (A8) we only need to observe the subset of confounders that cause heterogeneous mediator effects but not those that cause homogeneous mediator effects. Assume that the vector of confounders  $X$  is composed of a vector of covariates  $X_{het}$  that causes heterogeneity in the mediator effect and a vector of covariates  $X_{hom}$  that does not cause heterogeneity in the mediator effect. If the sequential ignorability assumption (A4a) holds, then we observe all relevant confounders, i.e. we observe  $X_i = X_{i,hhet} \cup X_{i,hom}$ . Under the assumption of conditional homogeneity of the mediator effect (A8) we observe  $X_{i,hhet} \subset X_i$  and have

$$\begin{aligned}E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = (m, 1 - m)] &= E[E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = (m, 1 - m), X_{i,hhet} = x_{het}]] \\ &\stackrel{A8}{=} E[E[Y_i(t, 1) - Y_i(t, 0) | X_{i,hhet} = x_{het}]] \\ &= E[E[Y_i(t, 1) - Y_i(t, 0) | X_{i,hhet} = x_{het}, X_{i,hom} = x_{hom}]]\end{aligned}$$

for  $t, m \in \{0, 1\}$ . For identification based on assumption (A8) we need to observe  $X_{i,hhet} \subset X_i$ . If we instead observe the richer set  $X_i = X_{i,hhet} \cup X_{i,hom}$  as under the sequential ignorability assumption (A4a), then the assumption of conditional homogeneity of the mediator effect (A8) will hold automatically. Thus, validity of the sequential ignorability assumption (A4a) implies validity of the assumption of conditional



homogeneity of the mediator effect (A8). The reverse is not true, though, because

$$\begin{aligned}
E[Y_i(t, M(t'))] &= E[E[Y_i(t, m)|M(t') = m]] \\
&= E[E[E[Y_i(t, m)|M(t') = m, X_i = x]|M(t') = m]] \\
&= E[E[E[Y_i(t, m)|M(t') = m, X_{i,het} = x_{het}, X_{i,hom} = x_{hom}]]|M(t') = m]] \\
&\stackrel{A1, A2, A4a}{=} E[E[E[Y_i|T_i = t, M_i = m, D_i = d, X_{i,het} = x_{het}, X_{i,hom} = x_{hom}]]|M(t') = m]] \\
&\neq E[E[E[Y_i|T_i = t, M_i = m, D_i = d, X_{i,het} = x_{het}]]|M(t') = m]]
\end{aligned}$$

for  $t, t', m, d \in \{0, 1\}$ . Here, conditioning on  $X_{i,het}$  is not sufficient because potential outcome levels also depend on the confounders  $X_{i,hom}$ . Thus, if we only observe  $X_{i,het}$  but not  $X_{i,hom}$  the assumption of conditional homogeneity of the mediator effect (A8) is satisfied but not the sequential ignorability assumption (A4a). Intuitively, sequential ignorability (A4a) imposes conditional mean independence on potential outcome levels while assumption (A8) imposes conditional mean independence on differences in potential outcomes. If conditional mean independence holds for potential outcome levels then it automatically holds for the differences in potential outcomes but the reverse is not true. Of course, if all confounders cause heterogeneous mediator effects, then both assumptions are equivalent. However, since they are not testable, in a given applications assumption (A8) is always weaker in the sense that it allows for some unobserved confounders, namely those that do not cause heterogeneous mediator effects.

The fact that the assumption of conditional homogeneity of the mediator effect (A8) relaxes the sequential ignorability assumption (A4a) highlights the importance of this alternative identification. It is not sufficient, though, to rank approaches 1 and 3 for all designs. In the parallel design and the mediator randomization designs approach 3 imposes the consistency assumption (A5), which is not required for approach 1, though. Moreover, in the independent mediator randomization design approach 3 also requires the assumption of no causal interaction between treatment and mediator (A6) to hold, which is also not needed for approach 1. Thus, ranking approaches 1 and 3 in these designs is only possible if these additional assumptions hold. In contrast, the treatment equivalence design only requires the randomization assumptions and the assumption of conditional homogeneity of the mediator effect (A8) for approach 3. As a consequence, we can conclude directly that approach 3 makes weaker assumptions than approach 1 in this design. From a methodological point of view, this makes the treatment equivalence design a particularly interesting case to study. In the following we apply approaches 1 (sequential ignorability), 2 (homogeneous mediator effect) and 3 (conditionally homogeneous mediator effect) to such a design. A unique feature of our application is that we directly measure counterfactual projects,  $M(0)$ , for all individuals in treatment  $T = 1$ , which allows us to identify the direct and indirect effect without having to impose additional assumptions. This provides us with an assumption-free benchmark that we can use to test the validity of approaches 1 to 3 and to compare them with each other.

## 4 Application

### 4.1 The experiment

We apply approaches 1 to 3 to data from a laboratory experiment we conducted at the Busara Center of Behavioral Economics in Nairobi, Kenya. With the experiment we investigate whether solidarity, which is a crucial base for informal insurance arrangements in developing countries, is sensitive to the extent to which individuals can influence their risk exposure. Our design measures subjects' monetary transfers to a worse-off partner both in a setting where participants could either deliberately choose (treatment CHOICE) or were randomly assigned (treatment RANDOM,  $T = 1$ ) to a safe or a risky project. The results of this experiments have been reported in Wunsch and Strobl (2018).

The experiment corresponds to a between-subject treatment equivalence design. The mediator of interest is the project subjects face, where  $M = 0$  for the safe project and  $M = 1$  for the risky project. The treatment of interest is free choice of the project,  $T = 0$ , versus random assignment of projects,  $T = 1$ . Thus, randomization of the mediator is one of the treatments of interest. Risk exposure serves as a mediator because the CHOICE treatment changes two things at the same time compared to the RANDOM treatment. Firstly, it changes the process by which transfer recipients become needy from pure bad luck to self-selection into risk, which may reduce transfers to worse-off partners due to attributions of responsibility and which is the direct behavioral effect of interest. Secondly, it changes the distribution of projects from  $M(1)$  in RANDOM to  $M(0)$  in CHOICE, which can lead to differences in transfers across treatments if transfers depend on project within treatments. This is an indirect mechanical effect caused by differences in risk exposure across treatments that biases the average treatment effect away from the behavioral effect of interest. In the following we summarize the main features of the experiment and refer to Wunsch and Strobl (2018) for more details.

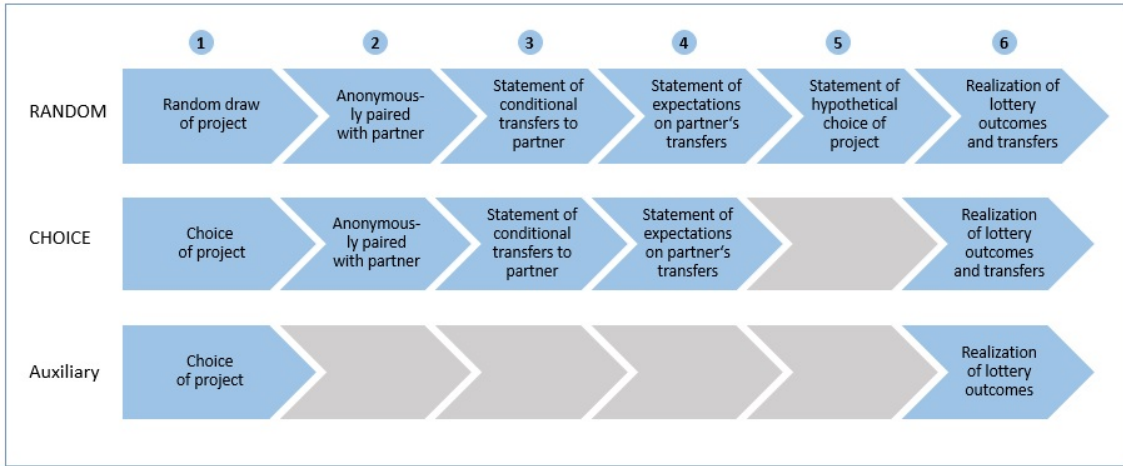
Figure 2 gives an overview on the sequence of steps in the game. At the beginning, two projects were presented to each subject: a safe option offering 500 KSh and a risky alternative yielding either 1000 or 0 KSh with equal probability. Depending on the treatment, subjects could either choose (treatment CHOICE) or were randomly assigned (treatment RANDOM) to one of these two options (step 1). After having chosen one project or being informed about the randomly received option, each subject was randomly and anonymously paired with another person in the room, who followed the same experimental procedure and was hence in the same treatment condition as the subject herself (step 2).<sup>8</sup> Using the strategy method, all subjects were then asked how much money they wanted to transfer to their matched partner in case of winning the 'high' payoff of their option, which is 1000 KSh for subjects holding the risky option and 500 KSh for individuals with the safe income. Hence, before revealing their own realized payoff as well as their partner's project and earnings, participants stated their gift for the two possible

---

<sup>8</sup>The subjects were informed about this step at the beginning of the game.

payoffs of their partner (i.e. 500 or 0 KSh) (step 3). Subsequent to the transfer statements subjects were asked which monetary amount they expected to receive from their partner for the two cases where the subject herself earned the ‘low’ payoff and the partner the ‘high’ payoff (i.e. 500 KSh or 1000 KSh) (step 4).<sup>9</sup> In the RANDOM treatment, the next step consisted of eliciting subjects’ preferred project. Here, we asked subjects which of the two projects they would have chosen had they had the possibility to choose (step 5). This way, we directly measure counterfactual projects under CHOICE,  $M_i(0)$ , for all subjects in RANDOM, which allows us to identify the direct and indirect effect without having to impose additional assumptions. At the end of the session, lottery outcomes of all participants were determined and transfers between the partners effected according to the actually realizing states (step 6). The stakes of the game represented considerable amounts for the mainly very poor participants who reported an average daily income of 160 KSh ( $\sim 1.60$  USD).

Figure 2: Sequence of steps in the risk solidarity game



In order to address concerns about the way we elicit preferences regarding the safe and risky projects, we ran an auxiliary treatment (Auxiliary) with a third subject pool.<sup>10</sup> The sole task in this incentivized game was to choose between the safe and risky project, corresponding hence to the project choice task in the CHOICE treatment (step 1). The participants played this game, however, in full autarky, i.e. they were not paired with another individual and transfers were not possible. The payoffs of this game corresponded to the safe amount or the realizing lottery outcomes, respectively. With the auxiliary experiment we address two concerns. Firstly, we can assess whether real monetary consequences matter for stated preferences by comparing the choices made in this experiment to the stated preference in the RANDOM treatment. Secondly, we address the issue that subjects’ choices might be driven by the transfers they may have to make. We find no differences in choices and stated preference between the main and the auxiliary experiment implying that such strategic considerations do not matter.

<sup>9</sup>These two expectation questions were not incentivized.

<sup>10</sup>We ran 5 sessions of the auxiliary treatment in January 2018 where, in total, 111 subjects participated. In these sessions, the same two games (investment game and risk preference game) as in CHOICE and RANDOM were played before the auxiliary experiment.

For recruitment, subjects were randomly chosen from the Kibera subject pool registered at Busara and then invited by SMS. A precondition for being selected was an education level of at least primary school (8 years) to ensure some familiarity with numerical values as is necessary for our study. Using a between-subject design, the recruited persons were randomly assigned to one of the two treatments. The core experiment was run within 13 sessions in December 2017. Six sessions were conducted of the RANDOM treatment and seven of the CHOICE treatment. In total, 238 subjects participated in our study, thereof 120 in RANDOM and 118 in CHOICE. 33% of our subjects are male and 47% are married. On average, the participants are 31 years old and have a schooling level of 11 years.

Upon arrival, subjects were identified by fingerprint and randomly assigned to a computer station. The instructions were then read out in Swahili by a research assistant, while simultaneously, some corresponding illustrations and screenshots were displayed on the computer screens (see Appendix A.4 for an English version of the instructions, exemplarily for CHOICE).<sup>11</sup> For the entire experiment the z-Tree software code (Fischbacher, 2007) was programmed to enable an operation per touchscreen which eases the use for subjects with limited literacy or computer experience. Subsequently, some test questions verified the participants' comprehension of the game rules. In case of a wrong answer, the subject was blocked to proceed to the following question. A research assistant then unlocked the program and gave some clarifying explanations if needed. This guaranteed that all participants fully understood the games and did not simply answer the test questions by trial and error. After the comprehension test, the participants performed the actual experimental task. The experiment involved, firstly, a risk preference game which aimed at measuring subjects risk attitudes and, secondly, the risk solidarity game explained in detail in above. Importantly, the subjects completed the decisions in these two games without learning the realized payoff in the precedent game. Moreover, after randomly determining the game payoffs at the end of the experiment, only the result of one randomly selected game was relevant for real payment. These two design features avoid that results are biased due to any strategic behavior, expectation forming or income effects across games.

At the end of the session, participants completed a questionnaire covering important individual and household characteristics. After the session, subjects received 200 KSh in cash as show-up fee which compensated mainly for the travel costs to the center. Moreover, subjects earned a minimum of 250 KSh in the experiment in order to guarantee an appropriate compensation for the time spent. However, participants were not informed about this minimum compensation before the end of the game. In total, average earnings amounted to 447 KSh per person. They were transferred cashless to the respondents' MPesa accounts.<sup>12</sup>

---

<sup>11</sup>All verbal explanations of the research assistant were made in Swahili whereas information on the computer screens was written in English. This combination has proven to be useful for facilitating comprehension (Haushofer et al., 2014).

<sup>12</sup>MPesa is a mobile-phone based money transfer service. It allows to deposit, withdraw and transfer money in a easy and safe manner with help of a cell phone. Its use is very widespread in Nairobi slums where around 90% of the residents have access to this service (Haushofer et al., 2014).

## 4.2 Main experimental results

Table 3 summarizes the main results from the experiment as presented in Wunsch and Strobl (2018) which serve as the starting point for our analysis. The upper part of the table contains the results for the full sample, the lower part the results for the sample that is restricted to the common support for risk-taking under CHOICE. Column (2) displays risk taking by treatment and its difference. In the RANDOM group, the randomization created equal proportions of safe and risky project holders as intended (line 2). When being able to choose the project freely in the CHOICE group, however, only a minority of the subjects preferred the risky lottery (line 1). Specifically, we observe 30 percentage points fewer persons with the risky project in CHOICE than in RANDOM, a difference which is highly statistically significant (line 4). Compared to RANDOM, average transfers are lower in CHOICE by 38 KSh in the full sample and by 37 KSh in the common support sample (column 3, lines 4 and 9). Although these differences are not statistically significant on conventional levels, the p-value are relatively small given the relatively small sample sizes with 17 and 20 percent, respectively (not reported).

To get a first idea about possible direct behavioral effects as opposed to indirect effects via differential risk-taking, we can hold risk-taking constant by looking at the effect of CHOICE on transfers conditional on risk exposure in columns (4) and (5) of Table 3. This is what the mediation literature calls the average controlled direct effect (Pearl, 2001; Robins, 2003; Acharya et al., 2016). In lines (4) and (9) we report the conditional treatment effects that use actual projects, i.e.  $M_i = M_i(0)$  for CHOICE and  $M_i = M_i(1)$  for RANDOM. These are not causal effects as they do not correct for selection into projects in CHOICE. In contrast, we present causal effects in lines (5) and (10), because here we condition on preferred projects, i.e. on  $M_i = M_i(0)$  for both treatments. Focusing on the causal effects, we find that transfers of safety choosers do not differ by treatment (column 4), while transfers of risk takers are significantly higher by about 200 KSh in RANDOM compared to CHOICE (column 5). In contrast, the effects that do not account for project choice show no significant differences for both safety choosers and risk takers. This points to relatively strong selection effects that need to be corrected for with approaches 1 and 2.

Table 3: Main experimental results

Outcome Sample	Subjects	Risk taking ( $M$ )	Transfers in KSh ( $Y$ )		
	All	All	All	$M_i = 0$	$M_i = 1$
	(1)	(2)	(3)	(4)	(5)
Full sample					
(1) Mean outcome for $T_i = 0, M_i = M_i(0)$ (CHOICE)	118	.195	168.61	161.01	200.00
(2) Mean outcome for $T_i = 1, M_i = M_i(1)$ (RANDOM)	120	.500	206.83	140.67	273.00
(3) Mean outcome for $T_i = 1, M_i = M_i(0)$ (RANDOM)	65	.262	209.23	142.29	398.24
(4) Difference (2)-(1)		.305***	38.22	-20.34	73.00
(5) Difference (3)-(1)		.067	40.62	-18.72	198.24*
Sample within common support					
(6) Mean outcome for $T_i = 0, M_i = M_i(0)$ (CHOICE)	113	.204	168.55	160.51	200.00
(7) Mean outcome for $T_i = 1, M_i = M_i(1)$ (RANDOM)	110	.491	205.64	142.68	270.93
(8) Mean outcome for $T_i = 1, M_i = M_i(0)$ (RANDOM)	59	.254	214.41	145.00	418.00
(9) Difference (7)-(6)		.287***	37.09	-17.83	70.93
(10) Difference (8)-(6)		.051	45.86	-15.51	218.00*

Note: Statistical significance \*\*\*(1%), \*\*(5%), \*(10%).  $M_i = M_i(0)$  mean that actual and preferred project are equal.

The fact that we find different effects for safety choosers and risk takers is an important result because it directly violates the assumption of no causal interaction between treatment and mediator (A6). This assumption requires that the treatment effect does not depend on the value of the mediator, i.e. that it is the same for risk takers and safety choosers. It is not required for identification in our treatment equivalence design but important for identification in parallel designs and independent mediator randomization designs. Our case reveals that asymmetric responses to a treatment conditional on the mediator are not unlikely, which implies that the assumption of no causal interaction between treatment and mediator (A6) is indeed a strong one. The new identification strategy based on the assumption of conditional homogeneity of the mediator effect (approach 3) allows estimating direct and indirect effects in cases where this assumption is violated in parallel designs while still exploiting the exogenous variation in the mediator generated in the auxiliary experiment.

### 4.3 Estimation of direct and indirect effects

#### 4.3.1 Benchmark

In the experiment, we directly measure counterfactual projects under CHOICE,  $M_i(0)$ , for all subjects in RANDOM. As a result, we can identify  $E[Y_i(1, M_i(0))]$  using average observed transfers of subjects assigned to and preferring the safe and, respectively, the risky project in RANDOM:

$$E[Y_i(1, M_i(0))] = E[Y_i|T_i = 1, M_i = M_i(0) = 0]Pr(M_i(0) = 0) + E[Y_i|T_i = 1, M_i = M_i(0) = 1]Pr(M_i(0) = 1).$$

This allows us to identify  $\delta_0$  and  $\gamma_1$  without having to impose additional assumptions given that we correctly measure  $M_i(0)$  in RANDOM. In Wunsch and Strobl (2018) we test extensively whether this is the case and find strong support for this claim. We use the estimated direct and indirect effects from the benchmark to test the validity of the identifying assumptions imposed in approaches 1-3 in our application.

#### 4.3.2 Approach 1: identification based on sequential ignorability

One advantage of identification based on sequential ignorability is that it allows estimating the counterfactuals  $E[Y_i(t, M_i(1-t))]$  for  $t \in \{0, 1\}$  directly. We use the identification result of Huber (2014):

$$E[Y_i(t, M_i(1-t))] = E\left[\frac{Y_i I\{T_i = t\}}{Pr(T_i = t|M_i = m, X_i = x)} \cdot \frac{Pr(T_i = 1-t|M_i = m, X_i = x)}{Pr(T_i = 1-t)}\right]$$

which results in the following probability weighting estimator

$$\hat{E}[Y_i(t, M_i(1-t))] = \frac{\sum_{i=1}^N w_i^t(m, x) Y_i I\{T_i = t\}}{\sum_{i=1}^N w_i^t(m, x) I\{T_i = t\}}$$

where

$$w_i^0(m, x) = \frac{\hat{p}_{T,i}(m, x)}{(1 - \hat{p}_{T,i}(m, x))\hat{p}_T}, \quad w_i^1(m, x) = \frac{1 - \hat{p}_{T,i}(m, x)}{\hat{p}_{T,i}(m, x)(1 - \hat{p}_T)}$$

and where  $\hat{p}_{T,i}(m, x)$  is an estimate of  $Pr(T_i = 1 | M_i = m, X_i = x)$  which handles self-selection into projects under CHOICE, and where  $\hat{p}_T$  is an estimate of the share of subjects in RANDOM,  $Pr(T_i = 1)$ . We estimate  $\hat{p}_{T,i}(m, x)$  parametrically using probit models, which results in a semi-parametric estimator of the direct effects.

Validity of the sequential ignorability assumption requires controlling for all factors that determine both subjects' project choice, and their transfers. Within the experiment, we collected basic socio-demographics as well as all information suggested to be important by theory and the empirical literature to render the sequential ignorability assumption (A4a) plausible. We expect that risk preference, which we measure with the risk preference game, is the most important determinant of project choice (see Appendix A.3 for all details on this game and the corresponding risk preference measure). Moreover, background risk theory (e.g. Gollier and Pratt, 1996) suggests that individuals reduce financial risk taking in the presence of other, even independent risks. Therefore, subjects' risk exposure in their real life might influence their decisions in the lab (Harrison et al., 2010). Also, individuals may give less in the presence of other risks because they want to preserve a certain capacity to cope with negative shocks with their own resources. We collected a broad range of variables reflecting exposure to the main sources of risk, such as income risk (occupation in paid employment, type of main occupation) and health and health expenditure risk (past and expected future health shocks, health insurance enrollment). Additionally, we have measures of the capacity to cope with negative shocks (wealth, household composition). Proxies for social capital and inequality aversion may also be relevant for predicting both project choice and transfers. Higher levels of trust and cooperation as well as inequality aversion in a society can encourage greater informal risk-sharing among community members and therefore provide better risk coping possibilities (Narayan and Pritchett, 1999). Moreover, higher social capital is found to promote financial risk-taking (Guiso et al., 2004). We observe five variables which are typically used to measure these factors (e.g. Giné et al., 2010; Karlan, 2005): trust, fairness, helpfulness and two measures of inequality aversion (see Appendix A.5 for the list of available variables and descriptive statistics for all subsamples).

The control variables we use to estimate the relevant propensity scores to correct for selection bias have been chosen as follows. Due to the relatively small number of observations we started with parsimonious specifications motivated by the descriptive evidence in Appendix A.5, theory, and previous empirical research. We then added covariates stepwise based on omitted variables tests (Lechner, 1995). Given that we include a measure of risk preferences and that many of the other variables are highly correlated, the number of covariates used is not too large with 6-7 variables. The exact specifications together with

the corresponding probit estimation results for the covariates are reported in Appendix A.6. Approach 1 also requires ensuring common support across treatments conditional on project choice. As can be seen from the descriptive statistics in Appendix A.5, we have lack of common support for two variables: subjects who have main occupation farming or who fall into the residual ethnicity category do not choose the risky project under CHOICE but are assigned to the risky project under RANDOM. We estimate all of our results with and without enforcing common support to show how this affects results, where the former excludes 6.3% of the sample.

#### 4.3.3 Approach 2: identification based on global effect homogeneity of $M$

Under global effect homogeneity we simply need to estimate the effect of being assigned the risky project in RANDOM,  $E[Y_i|M_i = 1, T_i = 1] - E[Y_i|M_i = 0, T_i = 1]$ , and multiply it with the effect of CHOICE on risk-taking behavior, i.e. with  $\theta_M = E[M_i|T_i = 1] - E[M_i|T_i = 0]$ . Both effects are directly identified from randomization of  $T$ , and of  $M$  in experiment 1. The results we present below in Table 4 do not correct for small sample covariate imbalances across the randomized samples because there are very few as can be seen in Appendix A.5. Correcting for these differences using inverse probability weighting do not change the results but reduce the precision of the estimates (available on request).

#### 4.3.4 Approach 3: identification based on conditional effect homogeneity of $M$

For approach 3 we implement the estimator derived in Section 3.4.2. On the one hand, it requires estimating the probability to choose the risky project under CHOICE as a function of variables that drive self-selection into the risky project and cause heterogeneous mediator effects,  $\hat{p}_{M,i}^0(x)$ , which is an estimate of  $Pr(M_i = 1|T_i = 0, X_i = x)$ . To determine the variables that need to be included, we systematically test for differences in transfers across projects within RANDOM for each available covariate, i.e. we test whether condition (C2) holds for the variables in our data. We regress transfers on an indicator for the risky project, the variable to be tested and an interaction term of this variable with the indicator for the risky project. The coefficient on this interaction term measures effect heterogeneity with respect to the tested variable. The corresponding estimation results are reported in Appendix A.7. From this procedure, we include all variables with statistically significant or nearly significant coefficients (p-values below 15%) and only leave out highly correlated variables. Moreover, we added two variables that turned out to be important in omitted variable tests. Appendix A.6 shows the included variables and probit estimation results. It is important to note that the set of control variables does not include the risk preference measure that is the most important control variable for approach 1. As a result, it will be particularly interesting whether we manage to match the results from the benchmark with approach 3 without including the most important selection variable, which would provide strong support for our claim that approach 3 makes weaker assumptions than approach 1. Approach 3 also requires a propensity score



to correct for small sample covariate imbalances across randomized projects in the RANDOM sample,  $\hat{p}_{M,i}^1(x)$ , which is an estimate of  $Pr(M_i = 1|T_i = t, X_i = x)$ . As can be seen in Appendix A.5, there are imbalances for three variables: inequality aversion 1, 8 years of schooling, and household had to forgo health care. The corresponding probit estimation results are reported in Appendix A.6.

### 4.3.5 Results

Table 4 reports all estimation results: the overall treatment effects (lines 1 and 2), the estimated direct and indirect effects for the benchmark (lines 3 and 4) and all three approaches (lines 4-15), and the differences between the direct effects obtained from the benchmark and from the different approaches (lines 16-18). From the benchmark we find that a large part of the overall treatment effect of 38 KSh is due to a direct behavioral response. The estimated direct effect corresponds to 62% of the total effect in the full sample and to 86% of the total effect in the common support sample. Due to our relatively small sample sizes, these effects are not statistically significant on conventional levels, though.

Table 4: Estimation results

	Full sample			Common support		
	Estimate	SE	P-value	Estimate	SE	P-value
<i>Overall effects</i>						
(1) $\theta_Y = E[Y_i T_i = 1] - E[Y_i T_i = 0]$	38.22	27.68	0.17	37.09	29.14	0.20
(2) $\theta_M = E[M_i T_i = 1] - E[M_i T_i = 0]$	0.31***	0.06	0.00	0.29***	0.06	0.00
<i>Benchmark: no assumptions</i>						
(3) Direct effect $\delta_0$	23.57	30.26	0.44	32.02	32.57	0.33
(4) Indirect effect $\gamma_1$	14.65	22.83	0.52	5.07	24.96	0.84
<i>Approach 1: sequential ignorability</i>						
(5) Direct effect $\delta_0$	25.53	40.47	0.53	27.69	44.64	0.54
(6) Indirect effect $\gamma_1$	12.70	39.81	0.75	9.39	43.12	0.83
(7) Direct effect $\delta_1$	-21.95	45.39	0.63	-26.96	57.04	0.64
(8) Indirect effect $\gamma_0$	60.18 <sup>+</sup>	39.13	0.12	64.04	51.15	0.21
(9) Difference $\delta_1 - \delta_0 = \gamma_0 - \gamma_1$	-47.48	51.92	0.36	-54.65	63.39	0.39
<i>Approach 2: global effect homogeneity of M</i>						
(10) $E[Y_i T_i = 1, M_i = 1] - E[Y_i T_i = 1, M_i = 0]$	132.33***	43.67	0.00	128.25***	46.79	0.01
(11) $E[M_i T_i = 1] - E[M_i T_i = 0]$	0.31***	0.06	0.00	0.29***	0.06	0.00
(12) Direct effect $\delta_0$	-2.15	24.80	0.93	0.23	25.90	0.99
(13) Indirect effect $\gamma_1$	40.37***	15.84	0.01	36.85**	16.00	0.02
<i>Approach 3: conditional effect homogeneity of M</i>						
(14) Direct effect $\delta_0$	22.62	35.35	0.52	27.53	40.55	0.50
(15) Indirect effect $\gamma_1$	15.60	25.09	0.53	9.56	30.69	0.76
<i>Comparison of <math>\delta_0</math> with benchmark:</i>						
(16) Approach 1 versus benchmark (5)-(3)	-1.96	38.62	0.96	4.32	43.05	0.92
(17) Approach 2 versus benchmark (12)-(3)	25.72	18.30	0.16	31.78 <sup>+</sup>	20.68	0.12
(18) Approach 3 versus benchmark (14)-(3)	0.95	25.37	0.97	4.49	31.27	0.89
Note: ***/**/*/+ indicates significance on the 1/5/10/15% level. Inference is based on bootstrapping with 1999 replications.						

With approach 1 (sequential ignorability) we obtain results that are very similar to the benchmark. The difference in the estimated direct effect  $\delta_0$  is only -1.96 KSh in the full sample and 4.32 KSh in the common support sample with p-values of 96% and 92%, respectively (see line 16). This suggests that we managed to remove selection bias with our control variables, which supports our claim that we collected information on all relevant confounders. Not least, this is probably owed to the fact that we include a measure of risk preferences. Given that we do not reject validity of the sequential ignorability assumption

in our data, approach 1 can be used to measure the direct effect for the distribution of projects as in RANDOM,  $\delta_1$ , and the corresponding indirect effect,  $\gamma_0$ . For these effects, we find rather different results, namely that the total effect is mainly due to the indirect effect via differential risk taking. The latter amounts to about 60 KSh and is close to statistical significance with a p-value of 12% in the full sample (see line 8). This provides further evidence for rejection of the assumption of no causal interaction between treatment and mediator (A6) in our data because this assumption requires  $\delta_0 = \delta_1$  and, correspondingly,  $\gamma_1 = \gamma_0$ . Despite a difference of about 50 KSh, the equality of the (in)direct effects is not rejected formally in line (9) but the p-values of 36% and 39% in the full and, respectively, common support sample suggest that this is rather due to lack of precision in our relatively small samples.

Approach 2 imposes the strong assumption of a homogenous mediator effect. Not surprisingly given that we reject condition (C2) for several observed variables, we find that this assumption is rejected by the data. In contrast to the benchmark, we find that almost all of the total effect is due to an indirect effect via differential risk taking. This effect is estimated to be about 40 KSh in both samples and highly statistically significant with p-values of at most 2% (see line 13). The difference to the benchmark is close to statistical significance on the 10% level and amounts to 25 KSh with a p-value of 16% in the full sample and to 31 KSh with a p-value of 12% in the common support sample (see line 17).

Relaxing the assumption of a homogenous mediator effect with approach 3 allows us to match the results obtained from the benchmark, which is expected given that we do not reject the sequential ignorability assumption. The differences to the benchmark are comparable to those of approach 1 with only 0.95 KSh at a p-value of 97% in the full sample and 4.49 KSh at a p-value of 89% in the common support sample (see line 18). What is interesting here is that we are able to match the benchmark without controlling for the risk preference measure that is the main driver of selection into the risky project. This shows that approach 3 indeed does not require including all confounders but only those that cause heterogeneity in the mediator effect. To explore this further, we re-estimate approach 1 with the control variables from approach 3 that do not include the risk preference measure. The results are displayed in Table 5. Although we still do not reject equality with the benchmark, the differences become 3 to 5 times larger and the p-values fall (see line 10). Moreover, we obtain considerably different results for  $\delta_1$  and  $\gamma_0$ . When we include the risk preference measure as control variable in approach 3 instead, the match with the benchmark improves further but the differences to the case without the risk preference measure are very small. Moreover, the standard error increases, which suggests that the improvement comes at the cost of reduced precision (see line 11). These checks confirm that the risk preference measure is an important control variable for approach 1 but not for approach 3.

Table 5: Results with different control variables

	Full sample			Common support		
	Estimate	SE	P-value	Estimate	SE	P-value
<i>Benchmark: no assumptions</i>						
(1) Direct effect $\delta_0$	23.57	30.26	0.44	32.02	32.57	0.33
(2) Indirect effect $\gamma_1$	14.65	22.83	0.52	5.07	24.96	0.84
<i>Approach 1: sequential ignorability with control variables from approach 3</i>						
(3) Direct effect $\delta_0$	33.75	48.27	0.48	45.15	46.29	0.33
(4) Indirect effect $\gamma_1$	4.47	45.00	0.92	-8.06	41.61	0.85
(5) Direct effect $\delta_1$	24.58	43.59	0.57	28.98	51.11	0.57
(6) Indirect effect $\gamma_0$	13.65	34.44	0.69	8.11	42.66	0.85
(7) Difference $\delta_1 - \delta_0 = \gamma_0 - \gamma_1$	-9.18	56.08	0.87	-16.17	58.73	0.78
<i>Approach 3: conditional effect homogeneity of <math>M</math> with risk preference measure as additional control variable</i>						
(8) Direct effect $\delta_0$	14.97	29.23	0.61	6.76	34.82	0.85
(9) Indirect effect	23.25	41.39	0.57	30.33	47.35	0.52
<i>Comparison of <math>\delta_0</math> with benchmark:</i>						
(10) Approach 1 versus benchmark (3)-(1)	-10.18	47.30	0.83	-13.13	45.23	0.77
(11) Approach 3 versus benchmark (8)-(1)	0.32	30.05	0.99	1.69	36.11	0.96
Note: Inference is based on bootstrapping with 1999 replications.						

## 5 Conclusion

This paper discusses identification and estimation of natural direct and indirect effects in different between-subject double randomization designs that randomize the mediator in an auxiliary experiment. We show that such designs allow identifying direct and indirect effects based on an assumption that is weaker than the assumption of sequential ignorability which is typically made in the literature. It allows for some unobserved confounders, namely those that do not cause heterogeneous mediator effects. As the mediator is randomized in the auxiliary experiment, extensive tests for effect heterogeneity, for example based on modern machine learning algorithms to uncover heterogeneity in the mediator effect (Tibshirani, 1996; Knaus et al., 2017), could be used in order to optimally select the control variables that render the assumption of conditional homogeneity of the mediator effect plausible in a given data set. If plausible, our approach can be used to assess the validity of the sequential ignorability assumption, which provides important insights on the control variables required to justify unconfoundedness for observational studies and single experiment designs. Additionally, our approach has the advantage that it may still be plausible in the parallel design if identification based on the assumption of no causal interaction of treatment and mediator as proposed by Imai et al. (2013) fails due to differential responses to the treatment for different mediator values.

We demonstrate estimation of direct and indirect effects based on different identification strategies using an experimental design that is particularly interesting because it allows testing alternative identification strategies without having to impose additional assumptions. The data we use stems from a laboratory experiment we conducted in Kenya that investigates whether solidarity, which is crucial for informal insurance arrangements in developing countries, is sensitive to the extent to which individuals can influence their risk exposure. A key feature of this experiment is that we directly observe counterfactual mediator values, which allows us to estimate the direct and indirect effects without having to

impose additional assumptions. We use this assumption-free benchmark to test alternative approaches that impose different identifying assumptions.

We find that validity of sequential ignorability cannot be rejected in our data, which implies that we managed to collect information on all relevant confounders within the experiment. This is an important input for non-experimental studies that analyze the relationship between risk-taking and solidarity. Furthermore, we find that identification based on the strong assumption of a homogeneous mediator effect is rejected by the data. However, when relaxing this assumption to conditional homogeneity of the mediator effect, we are able to match the results of the assumption-free benchmark. We also show that this approach is indeed less demanding than identification based on sequential ignorability because we match the benchmark without controlling for the key driver of selection. As a further important result we find strong asymmetries in the treatment effects conditional on mediator value, which indicates that the assumption of no causal interaction between the treatment and the mediator, which is crucial for identification in the parallel design as proposed by Imai et al. (2013) but not for our approach, is violated in our application. This shows that this assumption is indeed a strong one and that the approach we propose is an important alternative that can be applied in parallel designs even if this assumption is violated, which can be tested in the data.

## References

- Acharya, A., Blackwell, M., Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3), 512-529.
- Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444-455.
- Angrist, J. D., Fernandez-Val, I. (2013). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In: Acemoglu, D., Arellano, M., Dekel, E. (eds.). *Advances in Economics and Econometrics, Tenth World Congress, Volume III, Econometrics*, Chapter 11.
- Aronow, P.M., Carnegie, A. (2013). Beyond LATE: Estimation of the average treatment effect with an instrumental variable. *Political Analysis*, 21, 492-506
- Binswanger, H.P. (1980). Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, 62, 395-407.
- Busso, M., DiNardo, J., McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5), 885-897.
- Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 215-242.

- Dal Bó, P., Foster, A., Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review*, 100(5), 2205-2229.
- Deuchert, E., Huber, M., Schelker, M. (2018). Direct and indirect effects based on difference-in-differences with an application to political preferences following the Vietnam draft lottery. *Journal of Business and Economic Statistics*, DOI: 10.1080/07350015.2017.1419139.
- Deuchert, E., Wunsch, C. (2014). Evaluating nationwide health interventions: Malawi's insecticide-treated-net distribution programme. *Journal of the Royal Statistical Society: Series A*, 177(2), 523-552.
- Eckel, C., Grossman, P. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23(4), 281-295.
- Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10, 171-178.
- Flores, C. A., Flores-Lagunes, A. (2009). Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. IZA Discussion Paper No. 4237.
- Frölich, M., Huber, M. (2017). Direct and indirect treatment effects - causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5), 1645-1666.
- Giné, X., Jakiela, P., Karlan, D., Morduch, J. (2010). Microfinance games. *American Economic Journal: Applied Economics*, 2(3), 60-95.
- Gollier, C., Pratt, J.W. (1996). Risk vulnerability and the tempering effect of background risk. *Econometrica*, 64(5), 1109-1123.
- Guiso, L., Sapienza, P., Zingales, L. (2004). The role of social capital in financial development. *American Economic Review*, 94(3), 526-556.
- Harrison, G.W. Humphrey, S.J., Verschoor A. (2010). Choice under uncertainty: Evidence from Ethiopia, India and Uganda. *Economic Journal*, 120(543), 80-104.
- Haushofer, J., Collins, M., de Giusti, B., Njoroge, J.M., Odero, A., Onyango, C., Vancel, J., Jang, C, Kuruvilla, M.V., Hughes, C. (2014). A methodology for laboratory experiments in developing countries: Examples from the Busara Center. Unpublished manuscript.
- Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. In: *Proceedings of the American Statistical Association, Biometrics Section* (pp. 2401-2415). Alexandria, VA: American Statistical Association.

- Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6), 920-943.
- Huber, M., Lechner, M., Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, 175(1), 1-21.
- Huber, M., Lechner, M., Mellace, G. (2017). Why do tougher caseworkers increase employment? The role of programme assignment as a causal mechanism. *Review of Economics and Statistics*, 99(1), 180-183.
- Hsu, Y., Huber, M., Lai, T.C. (2017). Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting. Working Paper SES No. 482, University of Friburg.
- Imai, K., Keele, L., Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1), 51-71.
- Imai, K., Keele, L., Tingley, D., Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4), 765-789.
- Imai, K., Tingley, D., Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 5-51.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1), 4-29.
- Janevic, M.R., Janz, N.K., Dodge, J.A., Lin, X., Pan, W., Sinco, B.R., Clark, N.M.(2003). The role of choice in health education intervention trials: a review and case study. *Social Science and Medicine*, 56, 1581-1594.
- Joyce, T., Remler, D.K., Jaeger, D.A., Altindag, O., O'Connell, S.D., Crockett, S. (2017). On measuring and reducing selection bias with a quasi-doubly randomized preference trial. *Journal of Policy Analysis and Management*, 36(2), 438-459.
- Karlan, D. S. (2005). Using experimental economics to measure social capital and predict financial decisions. *American Economic Review*, 95(5), 1688-1699.
- Karlan, D.S., Zinman, J. (2009). Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica*, 77(6), 1993-2008.
- Knaus, M., Lechner, M., Strittmatter, A. (2017). Uncovering Treatment Effect Heterogeneity in Swiss Job Search Programs. IZA Discussion Paper No. 10961.
- Lechner, M. (1995). Some Specification Tests for Probit Models Estimated on Panel Data. *Journal of Business and Economic Statistics*, 13 (4), 475-488.

- Long, Q., Little, R.J., Lin, X. (2008). Causal inference in hybrid intervention trials involving treatment choice. *Journal of the American Statistical Association*, 103(482), 474-484.
- Marcus, S.M., Stuart, E.A., Wang, P., Shadish, W.R., Steiner, P.M. (2012). Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychological Methods*, 17(2), 244-254.
- Narayan, D., Pritchett, L. (1999). Cents and sociability: Household income and social capital in rural Tanzania. *Economic Development and Cultural Change*, 47(4), 871-897.
- Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2-3), 201-206.
- Park, S. (2015). Abstract: Identifying average causal mediation effects with multiple mediators in the presence of treatment noncompliance. *Multivariate Behavioral Research*, 50(1), 141-141.
- Pearl, J. (2001). Direct and indirect effects. In: Breese, J.S., Koller, D. (Eds.). *Proceedings of the 17th conference on uncertainty in artificial intelligence* (pp. 411-420). San Francisco: Morgan Kaufmann.
- Petersen, M. L., Sinisi, S. E., van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, 17(3), 276-284.
- Pirlott, A. G., MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, 66, 29-38.
- Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P., Hjort, N., Richardson, S. (Eds.). *Highly Structured Stochastic Systems* (pp. 70-81). Oxford: Oxford University Press.
- Robins, J. M., Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 143-155.
- Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2), 161-170.
- Rücker, G. (1989). A two-stage trial design for testing treatment, self-selection and treatment preference effects. *Statistics in Medicine*, 8(4), 477-485.
- Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. In: Sauermann, H. (Ed.). *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136-168). Tübingen: Mohr.

- Shadish, W.R., Clark, M.H., Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484), 1334-1356.
- Sloczynski, T. (2016). A general weighted average representation of the ordinary and two-stage least squares estimands. Unpublished manuscript.
- Strobl, R., Wunsch, C. (2018). Does voluntary risk taking affect solidarity? Experimental evidence from Kenya. CEPR Discussion Paper No. 12996.
- Tchetgen Tchetgen, E. J., Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40(3), 1816-1845.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.
- VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1), 18-26.
- Wunsch, C., Strobl, R. (2018). Risky Choices and Solidarity: Why Experimental Design Matters. CEPR Discussion Paper No. 12995.
- Yamamoto, T. (2014) Identification and estimation of causal mediation effects with treatment noncompliance. Unpublished Manuscript.
- Zheng, W., van der Laan, M. J. (2012). Targeted maximum likelihood estimation of natural direct effects. *International Journal of Biostatistics*, 8(1), 1-40.



## A Appendix

### A.1 Proof of Theorem 1

Consider the parallel design and the mediator randomization designs. We start by noting that

$$\begin{aligned}
\gamma_t^0 &= E[Y_i(t, M_i(1, 0), 0) - Y_i(t, M_i(0, 0), 0)] \\
&= E[Y_i(t, 0, 0) - Y_i(t, 0, 0) | M_i(1, 0) = 0, M_i(0, 0) = 0] Pr(M_i(1, 0) = 0, M_i(0, 0) = 0) \\
&\quad + E[Y_i(t, 1, 0) - Y_i(t, 1, 0) | M_i(1, 0) = 1, M_i(0, 0) = 1] Pr(M_i(1, 0) = 1, M_i(0, 0) = 1) \\
&\quad + E[Y_i(t, 1, 0) - Y_i(t, 0, 0) | M_i(1, 0) = 1, M_i(0, 0) = 0] Pr(M_i(1, 0) = 1, M_i(0, 0) = 0) \\
&\quad + E[Y_i(t, 0, 0) - Y_i(t, 1, 0) | M_i(1, 0) = 0, M_i(0, 0) = 1] Pr(M_i(1, 0) = 0, M_i(0, 0) = 1) \\
&= E[Y_i(t, 1, 0) - Y_i(t, 0, 0) | M_i(1, 0) = 1, M_i(0, 0) = 0] Pr(M_i(1, 0) = 1, M_i(0, 0) = 0) \\
&\quad + E[Y_i(t, 0, 0) - Y_i(t, 1, 0) | M_i(1, 0) = 0, M_i(0, 0) = 1] Pr(M_i(1, 0) = 0, M_i(0, 0) = 1)
\end{aligned}$$

for  $t \in \{0, 1\}$ . Under the consistency assumption (A5) this simplifies to  $\gamma_t^0 \stackrel{A5}{=} \gamma_t$  where

$$\begin{aligned}
\gamma_t &= E[Y_i(t, 1) - Y_i(t, 0) | M_i(1, 0) = 1, M_i(0, 0) = 0] Pr(M_i(1, 0) = 1, M_i(0, 0) = 0) \\
&\quad + E[Y_i(t, 0) - Y_i(t, 1) | M_i(1, 0) = 0, M_i(0, 0) = 1] Pr(M_i(1, 0) = 0, M_i(0, 0) = 1) \\
&\stackrel{A7}{=} E[Y_i(t, 1) - Y_i(t, 0)] [Pr(M_i(1, 0) = 1, M_i(0, 0) = 0) - Pr(M_i(1, 0) = 0, M_i(0, 0) = 1)] \\
&= E[Y_i(t, 1) - Y_i(t, 0)] E[M_i(1, 0) - M_i(0, 0)]
\end{aligned}$$

In the parallel design under the randomization assumptions (A1), (A2), (A3a) and (A3b) this becomes

$$\gamma_t = \{E[Y_i | T_i = t, M_i = 1, D_i = 1] - E[Y_i | T_i = t, M_i = 0, D_i = 1]\} \{E[M_i | T_i = 1, D_i = 0] - E[M_i | T_i = 0, D_i = 0]\}$$

for  $t \in \{0, 1\}$ . For the independent mediator randomization design we have

$$\begin{aligned}
\gamma_t &= E[Y_i(t, 1) - Y_i(t, 0)] E[M_i(1, 0) - M_i(0, 0)] \\
&\stackrel{A6}{=} E[E[Y_i(t, 1) - Y_i(t, 0) | T_i = t]] E[M_i(1, 0) - M_i(0, 0)] \\
&\stackrel{A1, A2, A3a}{=} E[E[Y_i | T_i = t, M_i = 1, D_i = 1] - E[Y_i | T_i = t, M_i = 0, D_i = 1]] E[M_i(1, 0) - M_i(0, 0)] \\
&\stackrel{A1, A2, A3a}{=} \{E[Y_i | M_i = 1, D_i = 1] - E[Y_i | M_i = 0, D_i = 1]\} \{E[M_i | T_i = 1, D_i = 0] - E[M_i | T_i = 0, D_i = 0]\}
\end{aligned}$$

for  $t \in \{0, 1\}$ . In the conditional mediator randomization design we have  $T_i = 0$  for all  $i$  with  $D_i = 0$  which implies that we can only identify

$$\gamma_0 = \{E[Y_i | M_i = 1, D_i = 1] - E[Y_i | M_i = 0, D_i = 1]\} \{E[M_i | T_i = 1, D_i = 0] - E[M_i | T_i = 0, D_i = 0]\}$$

under the randomization assumptions (A1), (A2, (A3a) and (A3b). In the treatment equivalence design with  $T_i = D_i$  we can identify

$$\begin{aligned}
\gamma_1 &= E[Y_i(1, M_i(1, 1), 1) - Y_i(1, M_i(0, 0), 1)] \\
&= E[Y_i(1, 1) - Y_i(1, 0) | M_i(1, 1) = 1, M_i(0, 0) = 0] Pr(M_i(1, 1) = 1, M_i(0, 0) = 0) \\
&\quad + E[Y_i(1, 0) - Y_i(1, 1) | M_i(1, 1) = 0, M_i(0, 0) = 1] Pr(M_i(1, 1) = 0, M_i(0, 0) = 1) \\
&\stackrel{A7}{=} E[Y_i(1, 1) - Y_i(1, 0)] [Pr(M_i(1, 1) = 1, M_i(0, 0) = 0) - Pr(M_i(1, 1) = 0, M_i(0, 0) = 1)] \\
&= E[Y_i(1, 1) - Y_i(1, 0)] E[M_i(1, 1) - M_i(0, 0)] \\
&= \{E[Y_i | M_i = 1, D_i = 1] - E[Y_i | M_i = 0, D_i = 1]\} \{E[M_i | T_i = 1] - E[M_i | T_i = 0]\}
\end{aligned}$$

where the last equality follows from the randomization assumptions (A1), (A2, (A3a) and (A3b).

For the direct effect we have  $\delta_{1-t} = \theta_Y^0 - \gamma_t$  where  $\theta_Y^0 = E[Y_i | T_i = 1] - E[Y_i | T_i = 0]$  in the treatment equivalence design and  $\theta_Y^0 = E[Y_i | T_i = 1, D_i = 0] - E[Y_i | T_i = 0, D_i = 0]$  in all other designs. ■

## A.2 Proof of Theorem 2

In the treatment equivalence design or under the consistency assumption (A5) in the other designs

$$\begin{aligned}
\gamma_t &= E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = (0, 1)] Pr(\tau_i = (0, 1)) - E[Y_i(t, 1) - Y_i(t, 0) | \tau_i = (1, 0)] Pr(\tau_i = (1, 0)) \\
&= \left[ \int E[Y_i(t, 1) - Y_i(t, 0) | X_i = x, \tau_i = (0, 1)] f_{X_i | \tau_i = (0, 1)}(x) dx \right] Pr(\tau_i = (0, 1)) \\
&\quad - \left[ \int E[Y_i(t, 1) - Y_i(t, 0) | X_i = x, \tau_i = (1, 0)] f_{X_i | \tau_i = (1, 0)}(x) dx \right] Pr(\tau_i = (1, 0)) \\
&\stackrel{A8}{=} \left[ \int E[Y_i(t, 1) - Y_i(t, 0) | X_i = x] f_{X_i | \tau_i = (0, 1)}(x) dx \right] Pr(\tau_i = (0, 1)) \\
&\quad - \left[ \int E[Y_i(t, 1) - Y_i(t, 0) | X_i = x] f_{X_i | \tau_i = (1, 0)}(x) dx \right] Pr(\tau_i = (1, 0))
\end{aligned}$$

for  $t \in \{0, 1\}$ . From Bayes' theorem it follows that

$$f_{X_i | \tau_i = (m, 1-m)}(x) = \frac{Pr(\tau_i = (m, 1-m) | X_i = x) f_{X_i}(x)}{Pr(\tau_i = (m, 1-m))}$$

for  $m \in \{0, 1\}$  which implies that

$$f_{X_i | \tau_i = (m, 1-m)}(x) Pr(\tau_i = (m, 1-m)) = Pr(\tau_i = (m, 1-m) | X_i = x) f_{X_i}(x).$$

Consequently,

$$\begin{aligned}
\gamma_t &= \int E[Y_i(t, 1) - Y_i(t, 0) | X_i = x] [Pr(\tau_i = (0, 1) | X_i = x) - Pr(\tau_i = (1, 0) | X_i = x)] f_{X_i}(x) \\
&= \int E[Y_i(t, 1) - Y_i(t, 0) | X_i = x] [Pr(M_i(1, 0) = 1 | X_i = x) - Pr(M_i(0, 0) = 1 | X_i = x)] f_{X_i}(x).
\end{aligned}$$

In the parallel design under the randomization assumptions (A1), (A2), (A3a) and (A3b) this becomes

$$\begin{aligned}\gamma_t &= \int \{E[Y_i|T_i = t, M_i = 1, D_i = 1, X_i = x] - E[Y_i|T_i = t, M_i = 0, D_i = 1, X_i = x]\} \\ &\quad \times \{Pr(M_i = 1|T_i = 1, D_i = 0, X_i = x) - Pr(M_i = 1|T_i = 0, D_i = 0, X_i = x)\} f_{X_i}(x).\end{aligned}$$

Let  $p_{M,i}^d(t, x) \equiv Pr(M_i = 1|T_i = t, D_i = d, X_i = x)$  and  $p_{M,i}^d(x) \equiv Pr(M_i = 1|D_i = d, X_i = x)$ . Using Bayes' rule we can rewrite

$$f_{X_i}(x) = \frac{Pr(M_i(t, d) = m) f_{X_i|M_i(t,d)=m}(x)}{Pr(M_i(t, d) = m)|X_i = x}$$

for  $t, d, m \in \{0, 1\}$ . This allows us to write

$$\gamma_t = E \left[ \left( \frac{Y_i M_i}{p_{M,i}^1(x)} - \frac{Y_i(1 - M_i)}{1 - p_{M,i}^1(x)} \right) (p_{M,i}^0(1, x) - p_{M,i}^0(0, x)) | T_i = t \right].$$

Following the same steps as in the proof of Theorem 1, we get for both mediator randomization designs

$$\begin{aligned}\gamma_t &= E \left[ \{E[Y_i|M_i = 1, D_i = 1, X_i = x] - E[Y_i|M_i = 0, D_i = 1, X_i = x]\} \right. \\ &\quad \times \{Pr(M_i = 1|T_i = 1, D_i = 0, X_i = x) - Pr(M_i = 1|T_i = 0, D_i = 0, X_i = x)\} \Big] \\ &= E \left[ \left( \frac{Y_i M_i}{p_{M,i}^1(x)} - \frac{Y_i(1 - M_i)}{1 - p_{M,i}^1(x)} \right) (p_{M,i}^0(1, x) - p_{M,i}^0(0, x)) | T_i = t \right].\end{aligned}$$

which is identified for  $t \in \{0, 1\}$  in the independent mediator randomization design and  $t = 0$  in the conditional mediator randomization design. For the treatment equivalence design we have

$$\begin{aligned}\gamma_1 &= E \left[ \{E[Y_i|M_i = 1, D_i = 1, X_i = x] - E[Y_i|M_i = 0, D_i = 1, X_i = x]\} \right. \\ &\quad \times \{Pr(M_i = 1|T_i = 1, X_i = x) - Pr(M_i = 1|T_i = 0, X_i = x)\} \Big] \\ &= E \left[ \{E[Y_i|M_i = 1, D_i = 1, X_i = x] - E[Y_i|M_i = 0, D_i = 1, X_i = x]\} \right. \\ &\quad \times \{Pr(M_i = 1|D_i = 1, X_i = x) - Pr(M_i = 1|D_i = 0, X_i = x)\} \Big] \\ &= E \left[ \left( \frac{Y_i M_i}{p_{M,i}^1(x)} - \frac{Y_i(1 - M_i)}{1 - p_{M,i}^1(x)} \right) (p_{M,i}^1(x) - p_{M,i}^0(x)) | T_i = 1 \right].\end{aligned}$$

For the direct effect we have  $\delta_{1-t} = \theta_Y^0 - \gamma_t$  where  $\theta_Y^0 = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$  in the treatment equivalence design and  $\theta_Y^0 = E[Y_i|T_i = 1, D_i = 0] - E[Y_i|T_i = 0, D_i = 0]$  in all other designs. ■

### A.3 The risk preference game

In the game, each subject was asked to choose one out of eight different lotteries (see Table 6, columns 2 to 4). The first alternative offers a certain amount of 320 Kenyan Shillings. The subsequent lotteries yield either a high (HEADS) or a low (TAILS) payoff with probability .5. While the first six lotteries are

increasing in expected values and variances of payoffs, the last lottery R has the same expected payoff as Q, but implies a higher variance. Hence, only risk-neutral or risk-loving subjects should choose this dominated gamble (Binswanger, 1980).

Table 6: Risk preference game: payoffs, expected values, risk and levels of risk aversion

Lottery number	Lottery	High payoff HEADS (p=.5)	Low payoff TAILS (p=.5)	Expected value	Standard deviation	Risk aversion range (CRRA) <sup>a</sup>	Fraction of subjects (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	K	320	320	320	0	2.46 to infinity	33.2
2	L	400	280	340	60	1.32 to 2.46	14.3
3	M	480	240	360	120	.81 to 1.32	9.2
4	N	560	200	380	180	.57 to .81	12.6
5	O	640	160	400	240	.44 to .57	2.5
6	P	720	120	420	300	.34 to .44	8.0
7	Q	800	80	440	360	0 to .34	11.3
8	R	880	0	440	440	-infinity to 0	8.8

Note: <sup>a</sup> As common in literature, we assume the individual's utility function  $u(x) = \frac{x^{1-\gamma}}{1-\gamma}$ , where  $\gamma$  is the CRRA parameter describing the degree of relative risk aversion. The intervals for the CRRA parameter were determined by computing  $\gamma$  where the expected utility from one option equals the expected utility from the next option, i.e. where the individual is indifferent between two neighboring lotteries.

Typically, the lottery numbers that subjects choose in ordered lottery designs (here: 1 to 8) are directly used as risk preference indicator (e.g. Eckel and Grossman, 2002). Strobl and Wunsch (2018) provide more details as well as a plausibility test of this measure.

#### A.4 Experimental instructions (exemplarily for CHOICE)

The entire experiment involved three games. Thereof, only two games are relevant for this study, with Game 2 corresponding to the risk preference game and Game 3 to the risk solidarity game. For the sake of simplicity, we therefore present a version of the original instructions shortened by the parts that are not relevant for this study.

## General instructions

Welcome and thank you for participating in our study. You are now taking part in an experiment on economic decision-making.

### Three Games:

In the following, you will play three short games, named [*Game 1*,] *Game 2* and *Game 3*. In each game, you will make one or several decisions. The result of your decision(s) will determine how much money you can finally earn in the respective game. We will explain later, how these three games work in detail.

### Payment:

However, please note that we will only pay you according to the result in one of the three games.

#### *How will we determine your payment?*

The computer will record what you have finally earned [in *Game 1*,] in *Game 2* and in *Game 3*. At the end of the experiment, the computer will randomly select [*Game 1*,] *Game 2* or *Game 3* with equal chance. We will pay you in shillings the final earnings you have made in this selected game. So, please remember that you will receive either your final earnings [from *Game 1* or] from *Game 2* or from *Game 3*, according to what game the computer will randomly select. Therefore, it is important to think carefully about the choice you make in each game.

### Test Questions:

Before each game starts, we will ask you to answer a few test questions to check if the rules of the games are clear to you. Please note that you will not get money for your answers and decisions in these test questions.

### Questionnaire:

After completing the three games, we will ask you to answer a few short questions about yourself and your household.

All your decisions and answers in this study will be kept confidential and only used for academic research purposes.

## Instructions for Game 2

[Game 2 is very similar to the game before. But please note that it is completely independent from Game 1]. Here is how Game 2 works.

### **Project Income:**

Assume that within your business, you have [again] a choice of 8 different income opportunities and you have to decide which one you want to realize. The table on your screen describes these income opportunities, named *Project K* to *R*:

[Screenshot 1]

Game 2			
Project	HEADS	TAILS	
K	320	320	OK
L	400	280	OK
M	480	240	OK
N	560	200	OK
O	640	160	OK
P	720	120	OK
Q	800	80	OK
R	880	0	OK

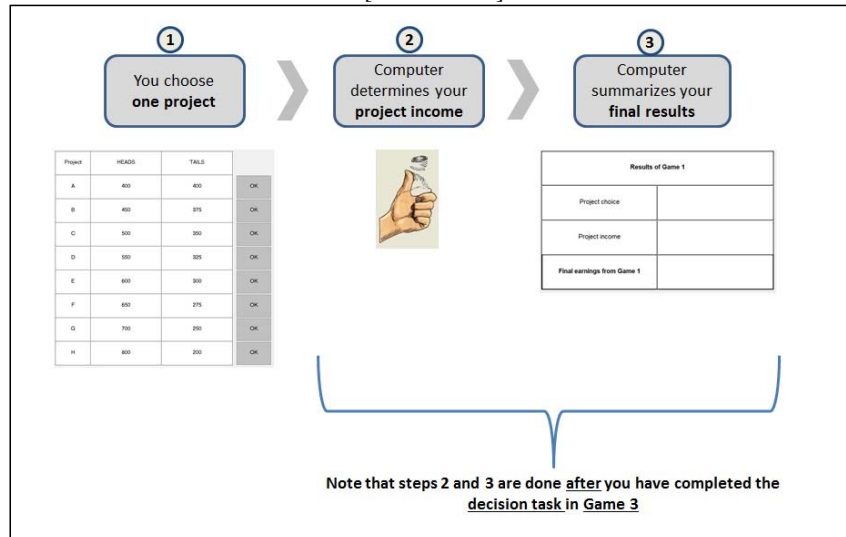
We will ask you to choose 1 out of the 8 projects. How much money you can earn from a project is [again] based on flipping a coin. [As in the game before,] the computer flips a coin after you have chosen your preferred project. If the coin lands on heads, you earn the amount given in the column “HEADS” in the row of your chosen project. If the coin lands on tails, you earn the amount given in the column “TAILS” in the row of your chosen project. Please choose the project that you prefer the most. There is no right or wrong answer.

### **Summary:**

The picture on your screen shows the sequence of events in Game 2.

Please note that steps 2 to 3 will be done after you have completed the decision task of GAME 3.

[Screenshot 2]



## Instructions for Game 3

In this game, you will make decisions that will determine your earnings and the earnings of another participant. Please note that Game 3 is completely independent of [Game 1 and] Game 2. Here is how Game 3 works.

### 1) Project Choice

In this game, you have a choice of 2 different income opportunities, named Project X and Y. The table on your screen describes these two projects.

[Screenshot 3]

Game 3:

Please choose one of the two projects.  
Please choose the one you prefer the most.  
There is no right or wrong answer.

Project	HEADS	TAILS
X	500	500
Y	1000	0

OK

OK

With each of these projects you can earn some income. We will ask you to choose 1 of the 2 projects. The amount of money you can earn from a project is again based on flipping a coin, as in Game [1 and] 2. If the coin lands on heads, you earn the amount in the column “HEADS” for your chosen project. If the coin lands on tails, you earn the amount in the column “TAILS” for your chosen project. Please choose the project that you prefer the most. There is no right or wrong answer.

## 2) Partner

After you have chosen your preferred project, the computer will randomly pair you with another person in this room. However, you will not know which person your partner is. His or her identity will be not revealed either during or after the game.

Your partner will also have already chosen either project X or Y. How much he/she will earn from the project is also determined by coin flip. Please note that another coin will be flipped for your partner, so that you both get individual results (i.e. heads or tails). Please also note that you will not know your partner’s project choice and project income until the end of Game 3.

## 3) Transfers

In this game, you can give some of your project income to your partner if you want to. Please note that you can give some of your income to your partner, but you do not have to. The amount that you decide to transfer to your partner will be deducted from your project income and added to your partner’s project income.

Just as you, your partner can give some of his/her income to you if he/she wants to, but he/she also does not have to. The amount that he/she decides to transfer to you will be deducted from his/her project income and added to your project income.



Please note that you both will decide how much you want to transfer to your partner before both of your project incomes are determined by coin flip. So, we will ask you both to decide in advance on the amount you wish to transfer for every possible combination of incomes you both might earn. The next 2 examples will explain the possible cases.

### ***Example 1 – You choose Project X***

Please look at your screen.

[Screenshot 4]

The screenshot shows a game interface titled "Game 3". At the top, it says "YOUR TRANSFERS:" with an icon of two people. Below this, there is a green box containing the text: "YOU will earn 500 KSh from Project X." followed by "How many shillings do YOU want to TRANSFER to your partner if your..". To the right of this box is a green box with "???". To the right of the green box is a text input field with a blue border and the text "Please enter any amount between 0 and 500 KSh:". Below the input field is a numeric keypad with buttons for 1, 2, 3, 4, 5, 6, 7, 8, 9, and 0. To the right of the keypad are two buttons: a red "Clear" button and a green "OK" button.

This screen appears, if you have chosen Project X. With Project X, you will earn 500 shillings, regardless of whether the coin lands on heads or on tails. We will ask you to decide how much you would like to transfer from your project income of 500 shillings to your partner. As the partner's income is not yet known, we will ask you to decide on your transfers for every possible amount that your partner might have earned with his/her chosen project.

Therefore, the first question (in green) ask what amount you would like to transfer from your project income of 500 shillings to your partner if your partner has also chosen Project X and earns 500 shillings. Please enter the amount that you would like to give to your partner by using the number pad. You can enter any amount between 0 and your full project income, that is 500 shillings in this example.

[Screenshot 5]

**Game 3**

**YOUR TRANSFERS:**

**YOU will earn 500 KSh from Project X.**  
 How many shillings do YOU want to TRANSFER to your partner  
 if your...

..partner has chosen Project X and earns 500	
..partner has chosen Project Y and earns 1000	
..partner has chosen Project Y and earns 0	???

Please enter any amount between 0 and 500 KSh:

1 2 3  
 4 5 6  
 7 8 9  
 0

Clear  
 OK

Similarly, the second and third questions ask what amount you would like to transfer to your partner if you earn 500 shillings and your partner has chosen Project Y and earns 1000 or 0 shillings. For each question, you can enter any amount between 0 and your full project income, that is 500 shillings. Your entered transfer amounts will appear in the small grey boxes (here on your screen, they are left empty).

Please note that later only one of the three possible partner's incomes will be realized, depending on which project your partner has chosen and what the result of the partner's coin flip is. The transfer amount that you have stipulated for exactly this realized partner's income will be deducted from your project income afterwards.

### Example 2 – You choose Project Y

[Screenshot 6]

Game 3

YOUR TRANSFERS:

**If YOU earn 1000 KSh from Project Y,**  
how many shillings do YOU want to TRANSFER to your partner  
if your...

..partner has chosen Project X and earns 500	
..partner has chosen Project Y and earns 1000	
..partner has chosen Project Y and earns 0	???

Please enter any amount between 0 and 1000 KSh:

1	2	3	Clear
4	5	6	
7	8	9	OK
0			

If you have chosen Project Y, you will earn 1000 shillings if the coin lands on heads and 0 shillings if the coin lands on tails. If you earn 0 shillings, you cannot make any transfers to your partner. If you earn 1000 shillings, you can transfer some money to your partner. So, we will ask you to decide how much you would like to transfer to your partner if you would earn 1000 shillings. As in Example 1, we will ask you to enter your transfer amounts for each of your partner's possible project incomes, that is 500, 1000 and 0 shillings. Again, you can enter any amount between 0 and your full project income, that is 1000 shillings in this case.

As already explained in Example 1, later only one of the three possible partner's incomes will be realized. The transfer amount that you have stipulated for exactly this realized partner's income will be deducted from your project income afterwards.

Please note that you and your partner make the transfer decisions simultaneously. Please also note that you will not know how much your partner has decided to give to you until the end of Game 3. Also, your partner will not know your transfer decisions until the end of Game 3.

#### 4) Expectation about the transfer you receive

After you have entered your three transfer decisions, we will ask you to estimate how much money your partner will transfer to you.

**Example 1 – You have chosen Project X**

[Screenshot 7]

Game 3

YOUR EXPECTATIONS:

**YOU will earn 500 KSh from Project X.**

How many shillings do you EXPECT that your PARTNER will give to YOU if your...

..partner has chosen Project X and earns 500

???

Please enter any amount between 0 and 500 KSh:

1 2 3  
4 5 6  
7 8 9  
0

Clear  
OK

This screen appears if you have chosen Project X. With Project X, you will earn 500 shillings, regardless of whether the coin lands on heads or on tails. The first question (in pink) asks how much money you expect to receive from your partner in the case that your partner has also chosen Project X and also earns 500 shillings. You can enter any amount between 0 and the full income of your partner, that is 500 shillings in this case.

[Screenshot 8]

Game 3

YOUR EXPECTATIONS:

**YOU will earn 500 KSh from Project X.**

How many shillings do you EXPECT that your PARTNER will give to YOU if your...

..partner has chosen Project X and earns 500

..partner has chosen Project Y and earns 1000

???

Please enter any amount between 0 and 1000 KSh:

1 2 3  
4 5 6  
7 8 9  
0

Clear  
OK

8 / 10

Similarly, the second question asks how much money you expect to receive from your partner in the case that you earn 500 shillings and your partner has chosen Project Y and earns 1000 shillings. You can enter any amount between 0 and the full income of your partner, that is 1000 shillings in this case. Please note that your partner CANNOT transfer money to you if he/she has chosen Project Y and earns 0 shillings, so we do not ask you about your expectations in this case.

**Example 2 – You have chosen Project Y**

[Screenshot 9]

The screenshot shows a game interface titled "Game 3". Under the heading "YOUR EXPECTATIONS:", there is a table with two columns. The first column lists scenarios, and the second column is for the user's expectation. The scenarios are:

Scenario	Expectation
...partner has chosen Project X and earns 500	
...partner has chosen Project Y and earns 1000	???

There are two red circles highlighting specific parts of the interface: one around the text "If YOU earn 0 KSh from Project Y." and another around the text "Please enter any amount between 0 and 1000 KSh:". To the right of the table is a numeric keypad with buttons for digits 1-9, 0, a "Clear" button, and an "OK" button.

Similarly, if you have chosen Project Y, you will earn either 1000 shillings or 0 shillings, depending on the result of your coin flip. We will, however, only ask you to enter how much money you expect to receive from your partner if YOU earn 0 shillings and YOUR PARTNER earns 500 shillings or 1000 shillings.

Please note that your partner will never be informed about your expectations. Also, you will never be informed about the expectations of your partner.

**5) Coin flip**

After you have entered the amounts that you expect to receive in transfers, the computer will determine your project income by flipping a coin. The computer will also determine your partner's project income by flipping another coin. The computer will now credit you and your partner with the transfer amounts that you each stipulated for each other for exactly the now realized incomes.

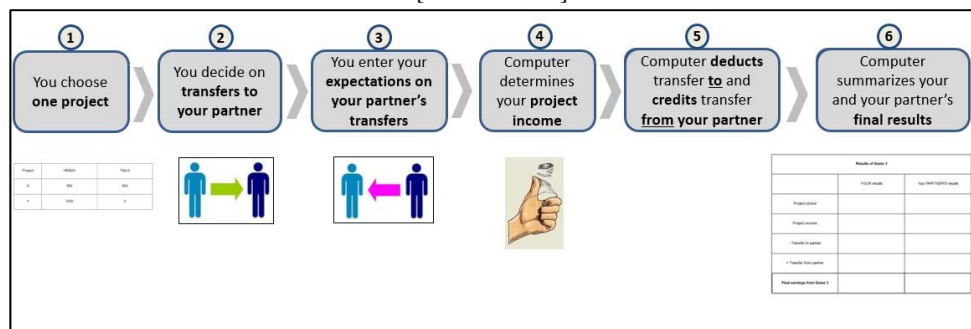
### 6) Final earnings of Game 3:

Your final earnings from Game 3 will be your project income MINUS the transfer that you made to your partner PLUS the transfer that your partner made to you.

### Summary:

The picture on your screen shows the sequence of events in Game 3.

[Screenshot 10]



## A.5 Full set of descriptive statistics

Table A1: Means of variables by treatment and project

	Random	All	Random (assigned)				Random (preferred)				Choice			
	(1)	Choice (2)	Difference (2)-(1)	Safe (3)	Risky (4)	Difference (4)-(3)	Safe (5)	Risky (6)	Difference (6)-(5)	Safe (7)	Risky (8)	Difference (8)-(7)	Difference (7)-(5)	Difference (8)-(6)
<b>A. Individual characteristics</b>														
Age	3.5	31.4	.9	3.1	3.8	.7	3.8	29.4	-1.4	31.2	32.1	.9	.4	2.6
Male	.33	.35	.02	.30	.35	.05	.32	.34	.03	.32	.48	.16	.00	.13
Schooling	11.5	11.2	-.3	11.2	11.9	.7	11.8	1.8	-1.0*	11.3	11.2	-.1	-.5	.4
8 years	.28	.27	.00	.35	.20	-.15*	.23	.41	.18*	.24	.39	.15	.01	-.02
8 < years < 12	.11	.11	.00	.12	.10	-.02	.11	.10	-.01	.13	.04	-.08	.02	-.06
12 <= years < 15	.46	.52	.06	.40	.52	.12	.48	.38	-.10	.54	.43	-.10	.05	.06
years >= 15	.16	.10	-.06	.13	.18	.05	.18	.10	-.07	.09	.13	.04	-.08	.03
Married	.45	.48	.03	.43	.47	.03	.43	.52	.09	.46	.57	.10	.03	.05
Household (HH) head	.64	.66	.02	.67	.62	-.05	.65	.62	-.03	.63	.78	.15	-.02	.16
Monthly income	4844	4897	53	4571	5117	546	4590	5641	1052	4569	6252	1684	-21	611
Religion (1=christian)	.84	.85	.01	.83	.85	.02	.87	.76	-.11	.86	.78	-.08	.00	.02
<i>Occupational status</i>														
Employed	.13	.14	.01	.15	.10	-.05	.09	.24	.15*	.13	.17	.05	.04	-.07
Self-employed	.19	.27	.08	.15	.23	.08	.20	.17	-.03	.25	.35	.10	.05	.18
Unemployed	.50	.45	-.05	.50	.50	.00	.52	.45	-.07	.46	.39	-.07	-.05	-.06
Other	.18	.14	-.04	.20	.17	-.03	.20	.14	-.06	.16	.09	-.07	-.04	-.05
<i>Ethnicity:</i>														
Kamba	.07	.05	-.02	.07	.07	.00	.08	.03	-.04	.05	.04	-.01	-.02	.01
Kikuyu	.07	.05	-.02	.07	.07	.00	.05	.10	.05	.04	.09	.04	-.01	-.02
Kisii	.13	.10	-.02	.13	.12	-.02	.12	.14	.02	.09	.13	.04	-.03	-.01
Luhya	.35	.36	.01	.33	.37	.03	.40	.21	-.19**	.37	.30	-.06	-.03	.10
Luo	.27	.31	.05	.28	.25	-.03	.25	.31	.06	.31	.35	.04	.05	.04
Nubian	.11	.10	-.01	.12	.10	-.02	.09	.17	.08	.11	.09	-.02	.02	-.09
Other	.02	.03	.01	.00	.03	.03	.01	.03	.02	.03	.00	-.03*	.02	-.03
<i>Health-related characteristics</i>														
Health problem	.33	.33	.01	.27	.38	.12	.33	.31	-.02	.32	.39	.08	-.01	.08
Chronic health problem	.13	.20	.08*	.15	.10	-.05	.09	.24	.15*	.18	.30	.13	.09*	.06
Visited health care provider	.43	.43	.00	.38	.48	.10	.45	.38	-.07	.43	.43	.00	-.02	.06
Health expenditures <sup>a</sup>	2497	1016	-1481	4242	752	-3490	3104	590	-2514	991	1115	124	-2113	525
Health expend.= 0	.57	.59	.03	.57	.57	.00	.56	.59	.03	.61	.52	-.09	.05	-.06
Enrolled in health insurance (HI)	.42	.44	.02	.48	.35	-.13	.36	.59	.22**	.44	.43	-.01	.08	-.15
Enrolled in other insurance	.08	.09	.01	.10	.07	-.03	.05	.17	.12	.09	.09	-.01	.04	-.09
<i>Social preferences</i>														
Inequality aversion 1 (disadv.) <sup>b</sup>	.18	.20	.03	.23	.12	-.12*	.15	.24	.09	.19	.26	.07	.04	.02
Inequality aversion 2 (adv.) <sup>b</sup>	.24	.32	.08	.30	.18	-.12	.23	.28	.05	.31	.39	.09	.07	.12
Fairness	.32	.34	.02	.32	.32	.00	.33	.28	-.05	.35	.30	-.04	.02	.03

Table A1: Means of variables by treatment and project (*continued*)

	All		Random (assigned)				Random (preferred)				Choice			
	Random	Choice	Difference	Safe	Risky	Difference	Safe	Risky	Difference	Safe	Risky	Difference	Difference	Difference
	(1)	(2)	(2)-(1)	(3)	(4)	(4)-(3)	(5)	(6)	(6)-(5)	(7)	(8)	(8)-(7)	(7)-(5)	(8)-(6)
Trust	.13	.19	.07	.15	.10	-.05	.14	.07	-.07	.21	.13	-.08	.07	.06
Helpfulness	.30	.31	.01	.30	.30	.00	.35	.14	-.21***	.32	.26	-.05	-.04	.12
GSS Index <sup>c</sup>	.74	.84	.10	.77	.72	-.05	.82	.48	-.34**	.87	.70	-.18	.05	.21
<b>B. Household characteristics</b>														
No. of adults	2.71	2.60	-.11	2.90	2.52	-.38	2.59	3.07	.48	2.63	2.48	-.15	.04	-.59
No. of children	2.49	2.44	-.05	2.52	2.47	-.05	2.68	1.90	-.78	2.07	3.96	1.88	-.61	2.06
Monthly per capita (p.c.) income	2072	2958	886*	1929	2215	286	1904	2600	696	2954	2974	20	1050*	374
No. of other earners	.63	.73	.10	.68	.57	-.12	.57	.79	.22	.82	.35	-.47***	.25*	-.45**
No. of dependent HH members	3.48	2.96	-.52	3.43	3.52	.08	3.65	2.93	-.72	2.88	3.26	.38	-.76	.33
HH is in wealth index quartile <sup>d</sup>														
Poorest quartile	.13	.17	.04	.12	.15	.03	.14	.10	-.04	.18	.13	-.05	.04	.03
Poorer quartile	.36	.44	.08	.30	.42	.12	.38	.28	-.11	.44	.43	-.01	.06	.16
Richer quartile	.27	.22	-.05	.30	.23	-.07	.26	.28	.01	.21	.26	.05	-.05	-.01
Richest quartile	.24	.17	-.07	.28	.20	-.08	.21	.34	.14	.17	.17	.01	-.04	-.17
<i>Health-related characteristics</i>														
Health expenditures (p.c.)	1634	692	-.941	2621	647	-.1975	1596	1753	157	678	751	73	-.918	-1001
Expected future health shock <sup>e</sup>	4.11	4.65	.54	4.27	3.95	-.32	4.18	3.90	-.28	4.43	5.57	1.13	.26	1.67
Foregone health care	.43	.51	.08	.35	.50	.15*	.44	.38	-.06	.52	.48	-.04	.08	.10
Prop. of HH members enrolled in HI	.25	.25	.00	.28	.21	-.08	.21	.38	.17*	.25	.24	-.02	.05	-.14
<b>C. Experimental outcomes</b>														
Risk preference <sup>f</sup>	3.42	3.59	.18	3.47	3.37	-.10	3.07	4.52	1.45***	2.99	6.09	3.10***	-.08	1.57**
Understanding of instructions <sup>g</sup>	1.22	1.23	.01	1.21	1.24	.03	1.21	1.26	.05	1.22	1.28	.07*	.01	.03
Observations	120	118		60	60		91	29		95	23			

Note: Statistically significant mean differences are marked as follows: \* $p < .10$ , \*\* $p < .05$ , \*\*\* $p < .01$ .; <sup>a</sup>in the past 3 months; <sup>b</sup>Inequality aversion 1 (disadvantageous); Dummy which takes the value 1 if respondent thinks that others should not own much more than herself; Inequality aversion 2 (advantageous): dto. ...not own much less...; <sup>c</sup>No. of GSS questions positively answered; <sup>d</sup>The wealth index bases on the ownership of 11 household items (house, land, poultry, goats, sheep, cows/bullocks, refrigerator, radio, bicycle, motorcycle, car) and is constructed by using weights generated by principal component analysis; <sup>e</sup>Expected likelihood of unaffordable HH health expenditures within next year; <sup>f</sup>No. of lottery the subject has chosen out of 8 different lotteries with an increasing degree of riskiness, with 1 (=safe income) to 8 (=riskiest lottery); <sup>g</sup>Average number of trials needed to answer the comprehension test questions correctly.



## A.6 Probit estimates for all propensity scores

Table A2: Probit estimates for all propensity scores

Dependent variable Sample	Full sample				Common support			
	$T_i$ $M_i = 0$ (1)	$T_i$ $M_i = 1$ (2)	$M_i$ $T_i = 1$ (3)	$M_i$ $T_i = 0$ (4)	$T_i$ $M_i = 0$ (5)	$T_i$ $M_i = 1$ (6)	$M_i$ $T_i = 1$ (7)	$M_i$ $T_i = 0$ (8)
Risk preference	0.110**	-0.257***			0.110**	-0.256***		-0.492
Wealth quartile 1				-0.362				0.393
Wealth quartile 3	0.712***	-0.496		0.279	0.712***	-0.502		-0.124
Wealth quartile 4	1.024***	0.416		-0.213	1.024***	0.377		0.202
Owms house				0.286				
Income (p.c.)	-8.48e-05***	-8.50e-05*		6.48e-05	-8.48e-05***	-8.30e-05*		5.68e-05
Health expenditures (p.c.)			0.406				0.406	
Foregone health care		-0.604				-0.561		0.538
Chronic health problem		0.410				0.415		
Domestic work				0.147				0.141
Self-employed	-0.718**				-0.718**		-0.627*	
Inequality aversion 1			-0.627*					
Inequality aversion 2		-0.782**				-0.733*		
Trust	-0.563*				-0.563*			
Household head				0.550				0.624*
Household size				0.0800***				0.0810***
No. of other earners				-0.483**		0.0999		-0.465**
Schooling (8 years)			-0.530*				-0.530*	
Constant	-0.606***	2.227***	0.0384	-1.479***	-0.606***	2.151***	0.0384	-1.674***
Observations	146	77	110	113	146	77	110	113

Note: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Note: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

## A.7 Analysis of heterogeneous mediator effects

Table A3 reports results from the following regression of transfers  $Y_i$  on the mediator  $M_i$  (indicator for the risky project) in the RANDOM sample ( $T_i = 1$ ):

$$Y_i = \alpha + \theta M_i + \beta_v V_i + \theta_v M_i V_i + \varepsilon_i.$$

where  $V_i$  is a single observed variable. This regression tests whether the mediator effect varies with this variable. The coefficients reported in Table A3 correspond to  $\theta_v$ . Columns (1) and (3) report the results without further control variables. Columns (2) and (4) report the results for a specification that includes the variables with small sample imbalances across projects within the RANDOM treatment (chronic health problem, household income per capita, schooling dummies).

Table A3: Analysis of heterogeneous mediator effects

	Full sample				Common support			
	(1)		(2)		(3)		(4)	
	No covariates Coeff.	P-val.	With covariates Coeff.	P-val.	No covariates Coeff.	P-val.	With covariates Coeff.	P-val.
Age	-7.99	0.20	-8.33	0.19	-7.87	0.24	-8.08	0.24
Male	26.74	0.78	10.88	0.91	36.80	0.72	23.35	0.82
Household size	14.40	0.15	14.46	0.16	14.08	0.17	14.33	0.18
Owns land	63.48	0.58	53.97	0.66	32.91	0.79	32.66	0.80
Owns house	214.30	0.04	228.45	0.03	202.78	0.07	210.80	0.07
Schooling (years)	0.66	0.97	1.12	0.95	1.34	0.94	2.55	0.89
8 years	-25.19	0.80	-31.20	0.76	-36.49	0.75	-42.09	0.72
8 < years < 12	121.13	0.39	143.06	0.32	127.41	0.37	145.62	0.33
12 <= years < 15	-72.25	0.41	-71.88	0.43	-80.32	0.39	-80.90	0.41
years >= 15	72.77	0.55	73.68	0.55	79.01	0.53	81.85	0.52
Married	-90.84	0.30	-80.86	0.36	-108.65	0.24	-98.04	0.30
Household head	149.03	0.10	130.31	0.16	159.89	0.09	138.09	0.16
Income	0.01	0.42	0.01	0.39	0.01	0.33	0.01	0.31
<i>Occupational status</i>								
Employed	-144.71	0.27	-145.22	0.30	-102.73	0.48	-92.40	0.55
Self-employed	211.52	0.06	267.23	0.03	217.55	0.07	298.61	0.02
Unemployed	-114.00	0.19	-108.25	0.23	-151.43	0.10	-153.73	0.11
Other	46.32	0.68	20.75	0.86	52.58	0.65	19.99	0.87
Christian	42.12	0.73	37.66	0.76	93.58	0.47	90.70	0.50
<i>Ethnicity</i>								
Kamba	45.71	0.79	11.44	0.95	5.07	0.98	-14.82	0.94
Kikuyu	-203.39	0.24	-193.14	0.28	-200.60	0.26	-191.05	0.30
Kisii	65.21	0.62	105.66	0.45	77.60	0.58	122.51	0.41
Luhya	-57.33	0.53	-59.16	0.55	-60.21	0.54	-61.64	0.56
Luo	131.43	0.18	110.94	0.28	149.43	0.15	127.87	0.24
Nubian	-77.73	0.58	-66.24	0.65	-106.45	0.48	-96.57	0.53
Other	-23.79	0.89	2.36	0.99	0.00		0.00	
Health problem	193.93	0.04	207.54	0.03	203.56	0.04	213.17	0.04
Chronical health problem	264.12	0.05	297.05	0.03	257.04	0.07	288.61	0.05
Health care	16.14	0.86	10.69	0.91	7.07	0.94	1.48	0.99
Health expenditures	0.03	0.17	0.03	0.17	0.03	0.23	0.03	0.20
Health expend.= 0	-29.82	0.73	-26.56	0.77	-4.19	0.96	2.53	0.98
Health insurance	129.62	0.13	164.80	0.08	115.87	0.21	146.87	0.15
Other insurance	202.09	0.19	191.29	0.23	186.67	0.29	187.96	0.31
Inequality aversion 1	34.83	0.77	27.07	0.83	0.03	1.00	-15.88	0.90
Inequality aversion 2	-5.73	0.96	-1.03	0.99	9.52	0.94	-4.24	0.97
Fairness	115.97	0.21	111.14	0.25	75.64	0.45	69.25	0.51
Trust	93.31	0.49	77.58	0.58	97.58	0.50	90.80	0.54
Helpfulness	-4.92	0.96	-15.58	0.87	-31.11	0.76	-36.77	0.72
GSS Index	51.08	0.36	40.26	0.49	31.62	0.59	23.32	0.70
No. of adults	16.32	0.48	16.40	0.50	17.96	0.45	19.46	0.44

Table A3: Analysis of heterogeneous mediator effects (*continued*)

	Full sample				Common support			
	(1)		(2)		(3)		(4)	
	No covariates		With covariates		No covariates		With covariates	
	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.
No. of children	18.05	0.15	18.32	0.15	17.25	0.18	17.53	0.19
Income (p.c.)	0.00	0.96	0.00	0.92	0.00	0.91	0.00	0.95
No. of other earners	-53.05	0.33	-43.51	0.45	-45.62	0.42	-35.93	0.55
No. of dependents	10.99	0.37	10.50	0.40	27.29	0.17	30.99	0.14
Health expenditures (p.c.)	0.04	0.01	0.04	0.01	0.04	0.01	0.04	0.01
Expected health shock	-15.20	0.22	-15.38	0.23	-14.46	0.28	-14.72	0.29
Foregone health care	-33.34	0.71	-40.35	0.66	-39.22	0.68	-47.28	0.63
Health insurance (prop.)	95.96	0.38	120.43	0.30	110.91	0.36	126.26	0.32
Risk preference	24.44	0.15	21.70	0.21	24.84	0.16	22.11	0.23
Risk averse	-83.65	0.43	-46.16	0.68	-72.34	0.52	-32.71	0.78
Understanding instructions	43.98	0.89	73.55	0.82	3.90	0.99	3.21	0.99
Wealth quartile 1	-98.88	0.44	-98.44	0.45	-58.07	0.68	-55.39	0.70
Wealth quartile 2	-142.93	0.11	-161.52	0.08	-179.27	0.06	-199.57	0.04
Wealth quartile 3	101.42	0.31	136.77	0.19	97.04	0.36	127.92	0.25
Wealth quartile 4	166.28	0.10	157.83	0.13	199.89	0.07	195.08	0.08

Note: Dependent variable: transfers  $Y_i$ . Sample:  $T_i = 1$  (RANDOM).